



3 1761 10374378 7














Digitized by the Internet Archive  
in 2023 with funding from  
University of Toronto

<https://archive.org/details/31761103743787>



Catalogue 12-001

# Survey Methodology

A Journal of Statistics Canada

June 1991      Volume 17   Number 1











Statistics Canada  
Social Survey Methods Division

# Survey Methodology

A Journal of Statistics Canada

June 1991      Volume 17   Number 1

Published under the authority of the Minister  
of Industry, Science and Technology

© Minister of Supply and Services Canada 1991

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise without prior written permission of the Minister of Supply and Services Canada.

June 1991

Canada: \$35.00

United States: US\$42.00

Other Countries: US\$49.00

Catalogue 12-001

ISSN 0714-0045

Ottawa

# SURVEY METHODOLOGY

## A Journal of Statistics Canada

The Survey Methodology Journal is abstracted in The Survey Statistician and Statistical Theory and Methods Abstracts and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods.

### MANAGEMENT BOARD

<b>Chairman</b>	G.J. Brackstone	
<b>Members</b>	B.N. Chinnappa	R. Platek
	G.J.C. Hole	D. Roy
	C. Patrick	M.P. Singh
	F. Mayda (Production Manager)	

### EDITORIAL BOARD

**Editor** M.P. Singh, *Statistics Canada*

#### Associate Editors

B. Afonja, <i>United Nations</i>	D. Holt, <i>University of Southampton</i>
D.R. Bellhouse, <i>U. of Western Ontario</i>	G. Kalton, <i>University of Michigan</i>
D. Binder, <i>Statistics Canada</i>	J.N.K. Rao, <i>Carleton University</i>
E.B. Dagum, <i>Statistics Canada</i>	D.B. Rubin, <i>Harvard University</i>
J.-C. Deville, <i>INSEE</i>	I. Sande, <i>Bell Communications Research, U.S.A.</i>
D. Drew, <i>Statistics Canada</i>	C.E. Särndal, <i>University of Montreal</i>
W.A. Fuller, <i>Iowa State University</i>	W.L. Schaible, <i>U.S. Bureau of Labor Statistics</i>
J.F. Gentleman, <i>Statistics Canada</i>	F.J. Scheuren, <i>U.S. Internal Revenue Service</i>
M. Gonzalez, <i>U.S. Office of Management and Budget</i>	C.M. Suchindran, <i>University of North Carolina</i>
R.M. Groves, <i>U.S. Bureau of the Census</i>	J. Waksberg, <i>Westat Inc.</i>
	K.M. Wolter, <i>A.C. Nielsen, U.S.A.</i>

#### Assistant Editors

J. Gambino, L. Mach and A. Théberge, *Statistics Canada*

---

### EDITORIAL POLICY

The Survey Methodology Journal publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

#### Submission of Manuscripts

The Survey Methodology Journal is published twice a year. Authors are invited to submit their manuscripts in either of the two official languages, English or French to the Editor, Dr. M.P. Singh, Social Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6. Four nonreturnable copies of each manuscript prepared following the guidelines given in the Journal are requested.

#### Subscription Rates

The price of the Survey Methodology Journal (Catalogue No. 12-001) is \$35 per year in Canada, US \$42 in the United States, and US \$49 per year for other countries. Subscription order should be sent to Publication Sales, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6. A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, and the Statistical Society of Canada.



# SURVEY METHODOLOGY

A Journal of Statistics Canada  
Volume 17, Number 1, June 1991

## CONTENTS

In This Issue .....	1
<b>New Approaches to Data Collection and Capture</b>	
G.S. WERKING and R.L. CLAYTON Enhancing Data Quality Through the Use of Mixed Mode Collection .....	3
P.A. PHIPPS and A.R. TUPEK Assessing Measurement Errors in a Touchtone Recognition Survey .....	15
J.M. BRICK and J. WAKSBERG Avoiding Sequential Sampling with Random Digit Dialing .....	27
J.G. BETHLEHEM and W.J. KELLER The Blaise System for Integrated Survey Processing .....	43
J.D. DREW Research and Testing of Telephone Survey Methods at Statistics Canada .....	57
<hr/>	
D.R. BELLHOUSE Marginal and Approximate Conditional Likelihoods for Sampling on Successive Occasions .....	69
I. SCHIOPU-KRATINA and K.P. SRINATH Sample Rotation and Estimation in the Survey of Employment, Payroll and Hours .....	79
P.S. KOTT Estimating a System of Linear Equations with Survey Data .....	91
L. BIGGERI and U. TRIVELLATO The Evaluation of Errors in National Accounts Data: Provisional and Revised Estimates .....	99





## In This Issue

Conducting a survey using traditional telephone methods is relatively inexpensive and maintains direct, albeit remote, contact between interviewer and respondent. These two characteristics of telephone surveys – low cost and human contact – explain to a large extent why such surveys became popular. The extensive research that took place during the last decade is summarized in the book *Telephone Survey Methodology* edited by R. M. Groves *et al.* (1988) and in Volume 4, Number 4 of the *Journal of Official Statistics*. This issue's special section, devoted to data collection and capture, presents further developments, with the telephone playing a major role in both old and new ways.

In the first paper of the special section, Werking and Clayton look at research on telephone collection methods done during the past seven years at the U.S. Bureau of Labor Statistics (BLS). They show how the research has led to the decision to implement computer assisted telephone interviewing (CATI) and touchtone data entry (TDE) in the Current Employment Statistics Survey during 1991 and 1992. The authors also mention a new technology, voice recognition, as an alternative to TDE.

Phipps and Tupek discuss one particular area of the above research at the BLS, namely, measurement errors in TDE surveys. They conclude that entering extra or incorrect digits has the most impact, but that this can be reduced via editing, particularly when longitudinal data are available. They also suggest ways of improving TDE systems.

The Mitofsky-Waksberg random digit dialing (RDD) procedure is a commonly-used method for selecting households. Brick and Waksberg discuss an inefficient operational feature of the procedure and propose a modified method. They study its statistical properties and provide guidelines for choosing between the original and modified variants of the procedure.

Bethlehem and Keller discuss Blaise, an innovative software package developed at the Netherlands Bureau of Statistics. They show how Blaise is used as a tool to integrate various steps in survey processing, including data collection and data entry. As the use of portable computers for sample surveys increases, tools such as Blaise will become indispensable.

Like most statistical agencies throughout the world, Statistics Canada has studied increased and improved telephone usage for survey taking. Recent research and testing for household surveys is summarized by Drew in the last paper of the special section. The implications of the results for the redesign of the Labour Force Survey are also presented.

For sampling on successive occasions, a normal linear regression model can be used. Bellhouse obtains the marginal and conditional likelihoods for the correlation matrix of this model. Applications of these likelihood methods are given, both for the case of simple random sampling and for more complex designs.

Schiopu-Kratina and Srinath describe the methodology of the Survey of Employment, Payroll and Hours (SEPH) conducted by Statistics Canada. SEPH is a large monthly establishment survey which uses a rotating sample. The determination of monthly sample size is quite complex due to continuous changes in the population. Some possible simplifications of the design are mentioned.

Kott shows how some techniques from design-based sample survey theory, namely the inclusion of sampling weights and the mean square error (mse) estimator based on linearization, can be used in the estimation of a system of linear equations. He shows that the use of sampling weights may be preferable when the existence of missing regressors is likely. Furthermore, the mse estimator based on linearization is nearly unbiased for many error structures.

The need to make information available quickly, combined with the need for accurate estimates based on all available data, leads to a process of revision of the national accounts estimates. Biggeri and Trivellato review recent developments in the analysis of the reliability of national accounts estimates that are subsequently revised. An empirical analysis using data from Canada, Italy and the United States is also presented.

### The Editor

As we were going to press, we learned that Dr. M.N. Murthy, Director of the Applied Statistics Centre, Madras, India passed away on 2 April, 1991. Dr. Murthy made important contributions to survey methodology and was the author of a well-known textbook, *Sampling Theory and Methods*. During his career he worked at the Indian Statistical Institute and the UNO-sponsored Statistical Institute for Asia and the Pacific. He was a Fellow of the ASA, the Royal Statistical Society and the International Statistical Institute. We were fortunate to have him serve on our editorial board. He will be missed by his colleagues and former students around the world. He is survived by his wife Vyjayanthi and daughter Shashi.

## **Enhancing Data Quality Through the Use of Mixed Mode Collection**

**GEORGE S. WERKING and RICHARD L. CLAYTON<sup>1</sup>**

### **ABSTRACT**

A chronic problem in the preparation of time critical estimates is the significant limitations inherent in data collection by mail. To address this issue, the U.S. Bureau of Labor Statistics has conducted an extensive 7 year research effort into the use of computer assisted telephone interviewing (CATI) and the computer assisted self interviewing (CASI) methods of touchtone and voice recognition self-response. This paper will summarize some of the significant results of this research covering both performance and cost data. The paper will conclude with a discussion of a large scale implementation program of these techniques for a monthly sample of 350,000 establishments.

**KEY WORDS:** Employment statistics; Revisions; CATI; Touchtone collection; Voice collection; Cost analysis.

### **1. INTRODUCTION**

#### **1.1 Employment Statistics in the United States**

On the first Friday of each month, the U.S. Bureau of Labor Statistics releases data on the United States' employment situation for the previous month. On release day, the Commissioner of Labor Statistics appears before the Joint Economic Committee of Congress and provides a detailed analysis of the current month's data and trends; at the same time, the data are made available to the news media and the financial and business communities. This closely watched set of statistics is the earliest indicator available on the previous month's economic activity and is used as a major gauge of the health of the U.S. economy. The data in the release cover employment, hours and earnings by detailed industry which are derived from the Bureau's 350,000 unit monthly establishment survey – the Current Employment Statistics (CES) survey – along with labor force and unemployment data which are derived from the Bureau's 60,000 unit household survey – the Current Population Survey (CPS).

The establishment survey data have many important economic uses. Due to the CES survey's size and timeliness in conjunction with the importance of the basic payroll statistics which it collects, the CES monthly estimates are not only used as principal economic indicators by themselves but they are also included in the development of many of the Nation's other major economic indicators including: Personal Income for the Gross National Product, the Index of Leading Economic Indicators, the Index of Coincident Indicators, the Industrial Production Index, Real Earnings measures and Productivity measures. While the timeliness and accuracy of the CES statistics are essential in analyzing the current economic conditions in the United States, the CES survey has had to rely on a mail data collection process since its inception in the early 1900s. This collection process results initially in the publication of "preliminary" estimates for higher level aggregates using only the sample returns received to date followed by "final" estimates two months later which use the full sample. The process of producing both preliminary and final estimates for a given month periodically yields a

<sup>1</sup> G. Werking and R. Clayton, Monthly Industry Employment Statistics Division, Bureau of Labor Statistics, Room 2089, 441 G Street, N.W., Washington, D.C., USA, 20212.



substantial revision to the initial estimates. These revisions not only affect the basic CES statistics but also the other statistics which make use of CES estimates as input. To address this issue, the Bureau initiated a research program into automated telephone data collection approaches with the objective of substantially reducing the size and frequency of large revisions in the preliminary estimates.

This paper provides an overview of the Bureau's 7 year research program into automated telephone collection techniques and summarizes some of the most significant results. The following sections describe the CES survey process; discuss the research program evaluating Computer Assisted Telephone Interviewing (CATI), Touchtone Data Entry (TDE) and Voice Recognition (VR) data collection methods; detail some of the significant research results covering both performance and cost data; and conclude with a discussion of the large scale implementation program of these methods in the CES survey.

## 1.2 Current Employment Statistics Survey

The CES survey, with 350,000 units, is the largest monthly sample survey in the United States. It is conducted by the Bureau as a Federal-State cooperative program under which the Bureau specifies the survey's sample design and operational procedures while each State conducts all data collection and edit reconciliation activities. The Bureau produces and publishes extensive monthly industry detail at the 2, 3 and 4 digit industry levels for the Nation as a whole while each State produces monthly State and area (270 Metropolitan Statistical Areas) estimates.

The CES estimates are widely regarded as highly accurate economic statistics. Once each year, complete (or universe) employment counts for the previous year become available from the Unemployment Insurance tax records; these counts are used to annually benchmark (realign) the CES sample estimates to these universe counts. The annual benchmark process yields more accurate current monthly estimates along with providing an annual estimate of overall survey error. The average difference in the CES final sample estimate versus the complete universe count over the past 5 years is under 0.2% with 4 years in the 1980s when the difference was approximately zero. While the CES final monthly estimates are regarded as highly accurate relative to the universe counts; the preliminary monthly estimates, which are based on approximately 50% of the mail sample returns, have been periodically subject to large revisions when compared to the final estimate that is available 2 months later. Over the years, some improvement in reducing the size of the monthly revisions has been made; however, periodic large revisions have been viewed as a byproduct of conducting a large decentralized mail data collection process.

The decade of the 1980s brought about a number of changes for the CES program which would significantly alter the urgency and options for resolving the monthly revision issue. The 1980s created a far more quality conscious user constituency and while the CES products had not necessarily deteriorated, the CES users' expectations on quality and "fitness for use" had greatly increased. Much of this new way of thinking is directly attributable to the efforts of Deming, Juran and others on the subject of quality management. The 1980s also saw a much greater focus on the uses and the importance of the CES payroll statistics in assessing the current health of the U.S. economy, but with the rise in the use and the visibility of the CES statistics came a corresponding user frustration with monthly revisions. The 1980s also ushered in some dramatic new technological breakthroughs, most notably in microcomputers. This new technology offered survey agencies many new opportunities for improving data collection control and quality that included: Computer Assisted Telephone Interviewing, Touchtone Data Entry, Voice Recognition, Computer Assisted Personal Interviewing and FAX. Several of these methods would ultimately offer options to significantly improve timeliness and quality at an equivalent or reduced ongoing program cost.

The 1980s saw the Bureau shift from experimental research in the CES survey to full production testing of some of the most advanced state-of-the-art automated collection techniques then available, with major implementation of these techniques scheduled for 1991.

## 2. CES RESEARCH PROGRAM

### 2.1 Research Goals

In the early 1980s, the Bureau began an extensive 7 year research effort into the causes of late response and alternative collection methods which could significantly increase response rates for the preliminary estimates. The focus of the survey research centered around obtaining answers to three basic questions:

- Are data available at the establishment in time to respond by the publication deadline for the preliminary estimates?
- Are there data collection methods which can ensure an 80-90% response rate under these tight time constraints?
- Can the cost of these data collection methods be controlled at about the same level as the current mail collection costs?

At the conclusion of the research program, a mixed mode CATI/TDE collection approach emerged which satisfied the response rate and cost constraints for the survey. The following sections provide a brief description of these personal computer (PC) based data collection methods, the research tests, the response rate results and the cost analysis. Further details on these tests are documented in the research papers listed in the references. Additionally, recent results on measurement error for Touchtone collection are presented in a paper by Phipps and Tupek, this issue.

### 2.2 Data Collection Methods

The CES survey has a very limited data collection time period available to meet the preliminary estimates publication deadline. The CES survey's reference date is the payroll period containing the 12th of the month; thus, there are only 2 1/2 weeks available to collect, keypunch, edit, tabulate, validate and publish the data. In order to meet these tight time constraints, a collection method must be able to obtain the required data as soon as they become available within the establishment. The four data collection methods studied are described in turn below.

**Mail** – The CES questionnaire is a single page mail-shuttle form which provides space for the employer to record 12 months of data. The employer receives the questionnaire in the mail each month on or about the 12th of the month (*i.e.*, the survey reference date) and subsequently fills in the row of data items corresponding to the current month. There are five basic data items collected: all employees, women worker employment and production (or nonsupervisory) worker employment, hours and earnings. Once completed, the employer mails the form back to the State agency where it is keypunched and edited. The form is then filed so that it can be mailed back to the employer for the collection of the next month's report. As indicated earlier, this process currently yields a 50% response rate in the 2 1/2 weeks available for the preliminary estimates.

**Computer Assisted Telephone Interviewing** – Under CATI collection, the employer is mailed the CES questionnaire once at the beginning of the year and retains it for recording each month's data throughout the year. Each month as the payroll data become available, the employer fills



in the data items for that month and waits for the prearranged CATI call from the State agency. When the State agency calls, the data are collected under CATI, edited and a time for the next month's collection call is arranged.

**Touchtone Date Entry** – Under TDE reporting, the employer does the same activities as under CATI except instead of waiting for the State agency's CATI call, the employer now calls an 800 telephone number connected to the touchtone PC located at the State agency. The employer then touchtone enters the data items following the prompts in the automated CES interview. As each data item is entered by the employer, it is read back for respondent verification.

**Voice Recognition** – VR data reporting is identical to touchtone collection except the employer no longer needs to have a touchtone phone. The employer now reads the data as they appear on the form and the voice PC translates and reads back the data to the employer for verification. The VR system is speaker independent and accepts continuous speech; it recognizes the digits 0 through 9 and "yes" and "no".

### 2.3 Research Tests

The Bureau began developing a PC-based CATI system in 1983 for use in a two State test that began in 1984 (Figure 1). The CATI system developed by the University of California at Berkeley was selected for the test and was subsequently used throughout the research effort. A small random sample of 200 units was selected in each State and collection procedures and systems were refined over the next 7 years. The initial research tests were highly successful in the response rates they achieved and the tests were expanded to 9 States in 1986 and then to a total of 14 States in 1988. The composition of the test sample was also changed in 1986. Instead of selecting random samples of the full CES sample, the subsequent research tests focused only on random samples of habitually late CES respondents (*i.e.*, those units which had a response rate of under 20% for the preliminary estimates publication deadline). Thus, the success of the new collection methods of CATI and TDE was measured in terms of their ability to move samples of reporting units with a 0-20% preliminary estimates response rate to a stable ongoing 80-90% response rate. By the end of the CATI research phase in 1990, the Bureau was collecting over 5,000 units monthly under CATI and had conducted well over a quarter of a million CATI interviews.

While CATI was proving to be highly successful in improving response rates, it also became clear by 1985 that ongoing CATI collection would be more expensive than the existing mail collection. At this time, a separate path of research was begun on how to reduce the cost of CATI, while still maintaining the high monthly response rates which it was achieving. While improvements were made in reducing the length of time required for a CATI interview, it was a new alternative PC-based telephone reporting method which would offer dramatic reductions in the collection costs of CATI.

By 1985, many U.S. banks were operating a version of touchtone entry verification for check cashing at drive-in windows. The Bureau identified a PC-based touchtone reporting system suitable for survey research testing and by 1986 was conducting a small two State test of this technique for collecting data. TDE was not viewed as a direct replacement for mail nor as a competitive method to be tested against CATI. CATI's role was to take habitually late responders and turn them into timely responders through personal contact and an educational process, while TDE's role was to take these timely CATI responders and maintain their response rates at the same high level, but at a greatly reduced unit cost. Over the 5 years of data collection, TDE has also proven to be a very successful and reliable method of telephone data collection. The research phase for TDE is now also being concluded with over 5,000 units continuing to report monthly under TDE across 14 States; in total the Bureau has collected over 100,000 schedules using this new automated reporting method.

As a natural follow-up to TDE, the Bureau is currently conducting several small research tests of a new Voice Recognition reporting system. Preliminary results for VR reporting have replicated the same high monthly response rates achieved under TDE, but with the important advantage that respondents find VR reporting more natural and generally prefer it over TDE. At this time, the cost of the VR hardware is approximately 15 times that of TDE; however, within several years as the initial costs of VR drop, this collection method should become a viable replacement for TDE.

2.4 Research Results

Over the past 7 years, the Bureau has been able to establish that payroll data is available in most firms prior to the publication deadline for the preliminary estimates and that CATI collection has the ability to take traditionally late mail responders (*i.e.*, 0-20% response rate for preliminary estimates) and within 6 months turn them into timely responders with response rates of 82-84% (Figure 1). These response rates have been remarkably stable over the years as the CATI sample has been expanded from 400 units to 5,000 units and the number of participating States increased from 2 States to 14 States. The research results indicate that the data do exist at most establishments in time to meet the publication deadline and that CATI collection can raise mail response rates by 60-80% for these late respondents and this rate can be maintained in the targeted range of 80-90% over long periods of time. The principal limiting factor in a respondent's ability to make the publication deadline was found to be the length of the firm's pay period (Figure 3). Employer pay periods are generally weekly, biweekly, semi-monthly or monthly. Weekly and semi-monthly payrolls can almost always be collected in time for publication with biweekly pay periods available most of the time; however, most monthly payroll systems close out well after the publication deadline. Monthly payrolls have been one of the largest factors in limiting the CATI response rates to the 82-84% range.

Several other important results have come out of the CATI research. Under CATI, approximately 60% of the respondents will have their data available on the prearranged date for the first call with the remaining 40% using the first call as a prompt call (Figure 1). This rate has varied little across States or over the years of testing. A small test is scheduled to be conducted to see if an advance postcard notice to the respondent shortly before the prearranged CATI contact date will significantly limit the number of callbacks required.

		1984	1985	1986	1987	1988	1989	1990
Mail	Resp. Rates	47%	47%	48%	49%	49%	51%	52%
CATI	Units	400	400	2000	3000	5000	5000	5000
	Resp. Rates	83%	84%	82%	84%	83%	84%	82%
	% Call Back	44%	42%	40%	41%	42%	41%	41%
	Av. Minutes	5.6	5.6	5.0	4.8	4.4	3.5	3.8
TDE & Voice	Units				400	600	2000	5000
	Resp. Rates				78%	80%	84%	82%
	% Call Back				45%	45%	43%	40%
	Av. Minutes				1.8	1.8	1.7	1.7

Figure 1. Research Summary



The average time for a CATI interview depends on the number of items to be collected, the time efficiency of the interview instrument, and the experience of the data collector. The average time for a CATI call (Figure 1) was reduced by one-third as the CATI instrument was streamlined and interviewers became more experienced. Another very important concern in the testing was the effect of CATI on sample attrition. There was some concern that employers would not want to be constantly bothered by telephone contacts and would drop out of the program. However, the sample attrition rate for CATI was about one-third of that for mail with almost no loss of large reporters under CATI. In summary, CATI appears to have come close to maximizing the achievable response rate for the preliminary estimates while also enjoying broad support from the respondents.

Due to the increased cost associated with CATI collection, the Bureau initiated research into touchtone collection. During the 4 years of testing, TDE has demonstrated the ability to take timely CATI reporters having 82-84% response rates and maintain these high rates under completely automated TDE reporting (Figure 1). The importance of this result lies in the cost savings under TDE collection versus CATI collection. One of the major concerns for TDE collection was that, unlike CATI where respondent contacts are scheduled throughout the day, TDE respondents might tend to call during the same time period thus generating busy signals and require an excessive number of touchtone PCs to handle peak load reporting. Fortunately, this was not the case, and while the touchtone PCs are on-line 24 hours a day, most calls are relatively uniformly distributed between 8am and 5pm (Figure 4). TDE respondents tend to require the same proportion of prompt calls as CATI respondents – approximately 40%. Methods are currently also being tested to reduce the TDE prompting workload. One major advantage for the respondent is that TDE collection requires only one-half of the time of a CATI interview with the average TDE interview lasting only 1 minute and 45 seconds. Additionally, touchtone phones are widely available at most establishments; current estimates indicate that over 80% of employers could report under touchtone data collection. While TDE reporting offers many advantages to the survey agency, its strongest feature is respondent acceptance; respondent reaction to touchtone reporting has been very positive due to its speed and convenience for the respondent.

One general observation concerning the development of a CATI research program is that it is not critical which CATI hardware or software system is used during the research phase as long as it is reasonably flexible for change. The final results from testing may suggest very different CATI requirements for production implementation than those originally required for the research program. The most important and time consuming activity is the development and refinement of the methods and procedures for respondent contact. Once effective methods and procedures are developed, the requirements for the “right” system become more obvious.

## 2.5 Cost Analysis

With the performance testing and respondent acceptance for CATI and TDE proving to be highly successful, the final phase of research shifted to analyzing the transitional costs of CATI and the ongoing costs of TDE collection.

The major “labor” and “non-labor” cost categories were studied for mail, CATI, and TDE collection (Figure 2). The study looked not only at estimates of current cost, but also at projected costs over the next 10 years using the current rate of increase for the major cost items. Since CATI was to play only a 6 month transitional role (*i.e.*, moving late responding mail units to on-time responding CATI units) prior to conversion to ongoing TDE, the major focus of the cost analysis was on the cost tradeoffs between ongoing mail versus ongoing TDE data collection.

Cost Category	Mail	CATI	Self-response (TDE & VR)
<b>LABOR</b>			
mail out	↗		
mail return	↗		
data entry	↗	↗	
edit and edit reconciliation	↗	↗	↗
nonresponse followup			↗
<b>NON-LABOR</b>			
postage	↗		↗
telephones		↘	↘
microcomputers		↘	↘

Recent Annual Price Change Factors

Labor	+5.7%	ECI, State and Local Government
Postage	+5.0%	U.S. Postal Service
Telephone	-1.7%	CPI-U, Intrastate toll calls
Microcomputers	-19.5%	PPI Experimental Price Indexes (16 bit computer)

Figure 2. Data Collection Costs (arrows show direction of recent price change)

For the labor categories, the monthly mailout, mailback, check-in and forms control operations of mail were replaced by a single annual mailout-only operation under TDE, thus eliminating a large monthly clerical operation in the States. The batch keypunching, keypunch validation, and forms control operations under mail were completely eliminated under TDE, where the respondent touchtone enters the individual firm's data and validates each entry. Another major quality and cost-efficiency advantage of TDE was that procedures for telephone nonresponse follow-up became far more feasible under TDE than mail. Under TDE, an accurate up-to-date list can be generated of respondents who have not yet called in their data, this list can then be used to conduct brief telephone prompting calls. Under mail, telephone follow-up activities of "apparent" nonrespondents were awkward since the State staff did not know whether the respondent's form was not yet completed, currently in the mail, in the State check-in process or at keypunch; in addition, respondents who had recently sent their form tended to resent the additional reminder for an activity that they perceived as completed. Due to the voluntary nature of the program and the uncertainty of a respondent's response status, telephone prompting under mail was only used for critical (large employers) units.

There were no significant cost savings made for edit reconciliation as the number of edit failures under TDE remained at about the same level as under mail. This was also true for postcard reminders where the number of postcards used under mail collection for late respondents was approximately the same as the number used under TDE, where respondents received an "advance" postcard notice to touchtone their data by the due date.



In the non-labor categories, the cost of postage under mail (currently 58 cents per unit) is replaced by the cost of a telephone call and the amortized cost of the TDE machine (together currently 46 cents per unit). Postage is a continually increasing cost with an annual price increase of approximately 5% (Figure 2). The rising cost of postage is driven by annual labor cost increases (+ 5.7%) and by fuel costs (also generally increasing) with labor accounting for over 80% of total postage costs. In contrast, under TDE the cost of telephone calls has been decreasing in recent years (- 1.7%), along with the cost of microcomputers (- 19.5%).

Excluding the additional new requirement of full telephone nonresponse prompting for TDE, there are demonstrable cost savings in shifting from mail to TDE. Perhaps more importantly, under a 10 year projection of future costs of these two collection methods (Figure 5), these savings grow substantially. Attempts will be made to redirect future cost savings from TDE to help offset the full nonresponse prompting activities.

There are several major conclusions concerning TDE reporting which have emerged from the performance and cost analysis review. (1) The traditional view that mail is the least expensive collection option available to statistical agencies is no longer true. The major technological breakthroughs of the 1980s in automated telephone collection can not only reduce the collection cost below that of mail but can also improve timeliness and control over the collection process. Additionally, over the next 5-10 years, the cost of mail will become even less cost competitive with these high-technology/low-labor collection approaches due to the increasing labor and postage costs associated with mail. (2) The transition of respondents from mail to TDE appears to cause very little disruption to monthly reporting. In the Bureau's follow-up interviews with respondents who were converted to TDE, results have shown that respondents have very little trouble adapting to this new method of reporting. Virtually all respondents completed their first month TDE report accurately and without assistance, with many respondents commenting on the ease of reporting under TDE. (3) TDE can be viewed as a reliable replacement method for survey data collection; over the past 4 years of collection there have been no major equipment failure problems or disruptions of the collection process. Minor equipment problems have been easily resolved using a back-up PC when required. In addition, future back-up protection for the State TDE collection process will involve the use of a call forwarding option to reroute calls to a central site should major problems occur at the State.

### 3. IMPLEMENTATION

#### 3.1 Major Issues

By the end of 1989, the Bureau had completed a very successful research program and had sustained the high performance levels over 7 years. However, there is a significant difference between the completion of successful research and full scale implementation of new methods. While over 10,000 units were being actively collected under these new techniques, these units represented under 3% of the CES sample. Proposed collection changes for a monthly sample of 350,000 units which has been collected for well over half a century under a decentralized State mail collection environment requires not only a very strong demonstrable user need but also broad-based support at national, regional and State levels.

As it turned out, the user need had begun to change in the early 1980s. During the 1980s, the U.S. economy experienced the longest sustained peace-time growth period in its history, with over 19 million jobs created, and unemployment rates at their lowest levels since the early 1970s. By the mid 1980s, economic policy was firmly focused on establishing non-inflationary economic growth. The monthly CES employment growth and wage data were being closely

monitored for signs of wage-induced inflationary pressures resulting from strong job growth during a period of low unemployment. With this greatly increased use and visibility of the monthly data, came a corresponding user frustration with the periodic large revisions to the preliminary estimates.

While monthly revisions to the preliminary estimates had always been a part of the CES survey process and even though the size of these revisions had been reduced over the years, large revisions of over 100,000 in the preliminary monthly employment estimates of over-the-month change were now being viewed as unacceptable. These user demands for greater accuracy in the preliminary estimates would lead the Bureau to develop proposals for the implementation of automated CATI and TDE collection methods into the U.S. government's largest monthly survey. While user demand is critical, major changes of this magnitude could not be undertaken without full support at the State level where data collection actually occurs. One guideline which remained constant throughout the research program was that the collection system was ultimately the States' data collection system and therefore must be designed to integrate well into their survey environment and create as minimal an organizational impact as possible. To that end, the CATI and TDE systems remained open for change throughout the research program. As many State suggestions and requirements as possible were taken into account with each new release of the systems. The success of much of the development work can be credited to the resiliency and endurance of the 14 research States as they made constant recommendations for improvements in systems and procedures. In the end, the CATI interview instrument had moved from an awkward simulated household survey type interview approach to a fast and efficient "screens" and "windows" approach well suited for capturing and editing longitudinal economic data. Thus, at the conclusion of the research phase, the systems and procedures were well tested and refined across a wide range of States. This approach to testing brought with it a strong sense of confidence in the methods and the systems at the State level. This would prove to be essential for the Bureau's proposed quick production implementation timetable of these state-of-the-art collection methods.

### 3.2 Approach and Impact

The main focus of the implementation proposal was the control of revisions in the preliminary estimates. Over the past 5 years, approximately 40% of all revisions were over 50,000 with 13% of the revisions exceeding 100,000 (*i.e.*, large revisions) (Figure 6). The goal of the implementation study was to identify a minimum set of late responders which, if obtained by the publication deadline, would control the size of revisions to what was considered to be in an acceptable level (under 50,000 revision in the over-the-month employment change). While one obvious approach was to convert all 175,000 late respondents to the new collection methods, this approach was considered to be lengthy and costly. While there was a need to responsibly control the size of revisions (*i.e.*, not necessarily completely eliminate all revisions), there was also a corresponding need to resolve this problem in as timely a fashion as possible (*i.e.*, convert the smallest number of units necessary to control revisions to under the 50,000 level).

Establishment surveys, unlike household surveys, generally have differential weighting for individual units with very large units being "certainty" units in the sample design. In the CES sample design, units with 100+ employees make up only 20% of the sample (*i.e.*, 75,000 units), but account for over 83% of the unweighted sample employment. These units tend to have a much lower response rate for the preliminary estimates so that if the late respondents' employment trend differs from the early respondents, these units can create a substantial revision in the sample estimates. Revision impact studies were conducted to assess the affect of large employers 100+ on the preliminary estimates. To test the impact of large employers,



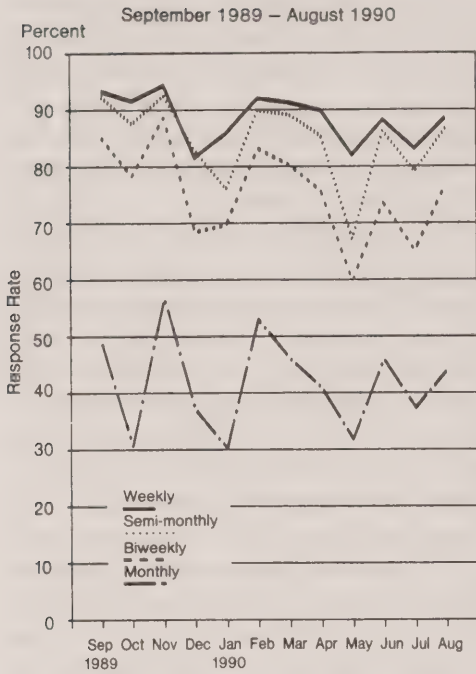


Figure 3. CES CATI First Closing Performance by Length of Pay Period

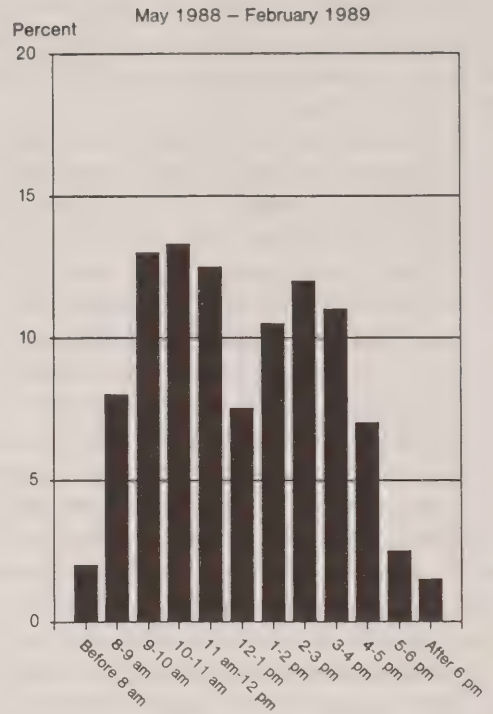


Figure 4. Touchtone Data Entry Distribution of TDE Calls by Time of Day

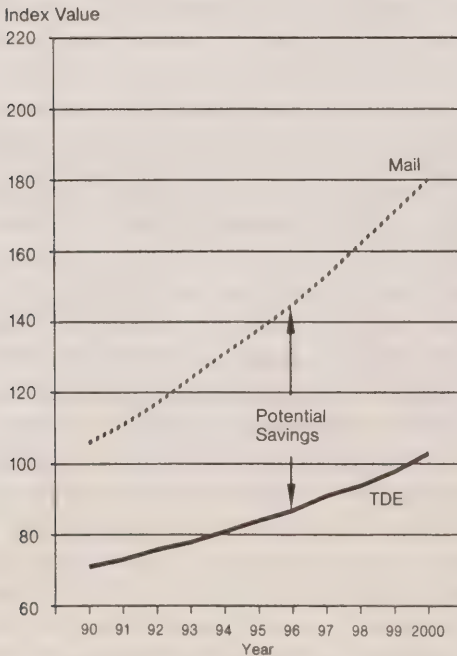


Figure 5. Estimated Unit Costs by Mode: 1990-2000

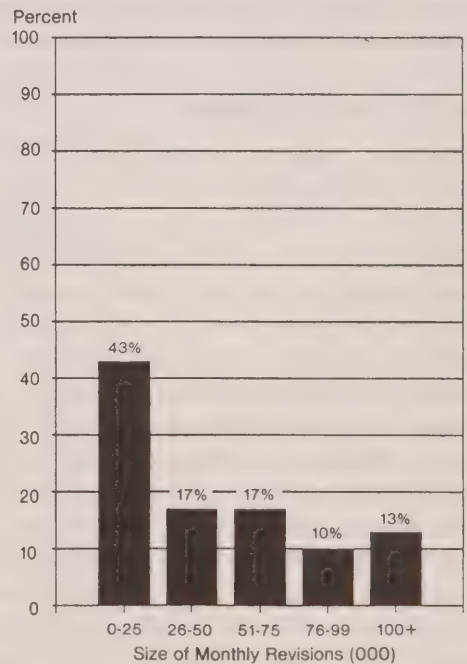


Figure 6. Distribution of Magnitude of Revisions (1985-1989)

late respondents in size class 100+ were included in the original sample used for the preliminary estimates and the estimates were recalculated. These new estimates were then compared to the original preliminary estimates to determine the impact of 100+ employers on revisions. The results indicated that from one-half to two-thirds of the revision was attributable to these units. These studies were repeated over several months with similar results. Applying these projected reduction rates in revision size to revisions over the past 5 years resulted in over 97% of all revisions being below the 50,000 level as compared to the current level of only 60%. This greatly reduced targeted sample size for conversion to CATI/TDE provided for an accelerated implementation time schedule consistent with controlling conversion costs to the minimum level necessary to protect against large revisions in the preliminary estimates. The Bureau will begin the implementation of CATI/TDE collection methods in 25 States in 1991 with planned expansion to all States in 1992. With the implementation of the new collection methods, the Bureau will be able to help resolve one of the most difficult and visible quality issues affecting the CES user community.

#### 4. SUMMARY

The decade of the 1980s has brought about many changes for survey agencies. Some of the changes can be viewed in terms of our accomplishments made over the decade while others are more subtle and need to be viewed in terms of the changes in the survey environment in which we operate.

The 1980s created a far more quality conscious user constituency which is quick to identify and point out our product limitations. While our products may not have deteriorated, our users' expectations on quality and "fitness for use" have greatly increased. This is an issue which we as statistical agencies must be able to respond to in order to maintain our credibility with the user community. The 1980s also ushered in dramatic new technological breakthroughs most notably in microcomputers. The new technology has offered survey agencies many new opportunities for improving data collection control and quality including: CATI, CAPI, TDE, VR and FAX. Some of these options offer improved quality and control at lower ongoing costs. The decade of the 1990s may well offer even greater opportunities for using technology to improve our data collection timeliness and quality at lower costs.

As we look at the status of our statistical programs, we often find very rigid environments. The data collection approach for our surveys often date back to their inception. Our data collection cost assumptions and cost studies are usually well outdated and often simplistic in approach. Since data collection generally represents the largest part of a survey's cost, it is usually well entrenched in the agency's organizational structure and can be quite difficult to restructure in order to accommodate large scale change. It is within this survey environment that we will face the major challenges and opportunities of the 1990s.

The challenge for statistical agencies in the 1990s will be threefold:

- to be responsive to the changing quality needs of our users;
- to attempt to have our research stay up with the rapid change of technology and automated data collection approaches; and perhaps more importantly
- to continue to find ways to incorporate successful research into our ongoing programs.

These challenges will determine the cost and quality competitiveness of our programs and our agencies in the future.



## REFERENCES

- CLAYTON, R.L., and HARRELL, L., Jr. (1989). Developing a cost model of alternative data collection methods: MAIL, CATI and TDE. *Proceedings of the Section on Survey Research Methods, American Statistical Association*.
- CLAYTON, R.L., and WINTER, D.L.S. (1990). Speech data entry: Results of the first test of voice recognition for data collection. *Proceedings of the Section on Survey Research Methods, American Statistical Association*.
- JURAN, J.M., GRZYNA, F.N., Jr., and BINGHAM, R.S., Jr., eds. (1979). *Quality Control Handbook*, Third Edition. McGraw Hill.
- GROVES, R.M.J., et al., eds. (1988). *Telephone Survey Methodology*. New York: John Wiley and Sons.
- OFFICE OF MANAGEMENT AND BUDGET (1988). *Quality in Establishment Surveys*. Statistical Policy Working Paper 15.
- OFFICE OF MANAGEMENT AND BUDGET (1990). *Computer Assisted Survey Information Collection*. Statistical Policy Working Paper 19.
- PONIKOWSKI, C., and MEILY, S. (1988). Use of touchtone recognition technology in establishment survey data collection. Presented at the First Annual Field Technologies Conference, St. Petersburg, Florida.
- WERKING, G.S., TUPEK, A.R., PONIKOWSKI, C., and ROSEN, R. (1986). A CATI feasibility study for a monthly establishment survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 639-643.
- WERKING, G.S., and TUPEK, A.R. (1987). Modernizing the Current Employment Statistics Program. *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, 122-130.
- WERKING, G., TUPEK, A., and CLAYTON, R. (1988). CATI and touchtone self-response applications for establishment surveys. *Journal of Official Statistics*, 349-362.

## Assessing Measurement Errors in a Touchtone Recognition Survey

POLLY A. PHIPPS and ALAN R. TUPEK<sup>1</sup>

### ABSTRACT

Electronic data collection utilizing touchtone recognition is in place for a monthly establishment survey at the Bureau of Labor Statistics. The Touchtone Data Entry (TDE) system features digitized phrases requesting respondents to answer questions using the numeric keypad of a touchtone telephone. TDE has substantial implications for lowering survey costs; many labor intensive activities are eliminated. However, little is known about measurement errors associated with this mode of data collection. This study assesses TDE mode error using three sources of data, which allow for analyses of errors associated with selected aspects of the human-machine interface. In addition, instrument design issues associated with mode error are addressed. We conclude by extending the implications of our findings to other surveys.

**KEY WORDS:** Mode of data collection; Human-machine interface; Computer-assisted self interviewing.

### 1. INTRODUCTION

The U.S. Bureau of Labor Statistics (BLS) issues monthly employment estimates for the United States from a survey of 350,000 business establishments. This survey, the Current Employment Statistics (CES) survey, provides one of the earliest monthly measures of U.S. economic health. However, the preliminary estimates from the survey are released with data from only about one-half of the business establishments in the survey. Revised estimates are produced two months after the initial press release. The low response rate for the initial press release can result in large revisions to the estimates. The BLS began investigating the use of automated collection techniques in 1983 to increase the timeliness of response and reduce the potential for large revisions.

The CES survey has traditionally been collected by mail through state employment security agencies. Research tests conducted between 1984 and 1986, involving the replacement of mail collection with computer-assisted telephone interviewing (CATI), have shown CATI to be an effective means for improving the timeliness of response (Werking, Tupek, Ponikowski and Rosen 1986). Average response rates under CATI collection have been between 85 and 90 percent for preliminary estimates, compared to 45 to 50 percent with mail collection. While CATI collection has been effective in improving the timeliness of response, the cost of full CATI collection in the CES survey cannot be absorbed within the survey's current budget. Research has been conducted since 1986 on touchtone data entry (TDE) to develop an alternative collection method with the performance gains of CATI, but with a lower unit cost (Ponikowski, Copeland and Meily 1989). Further discussion on the use of CATI and TDE collection for the CES survey is provided in the paper by Werking and Clayton, this issue.

Recent tests of the TDE system provide data on the timeliness of response, the cost of collection, and edit failure rates. BLS tests show that TDE collection with CATI back-up is as timely and effective as CATI collection (Werking, Tupek and Clayton 1988). In addition,

<sup>1</sup> Polly A. Phipps, Office of Employment and Unemployment Statistics, Bureau of Labour Statistics, Room 2821, 441 G Street N.W., Washington, D.C. 20212; Alan R. Tupek, Office of Employment and Unemployment Statistics, Bureau of Labor Statistics, Room 2919, 441 G Street N.W., Washington, D.C. 20212.

TDE lessens survey costs considerably. Current cost estimates indicate that the monthly unit costs for TDE are approximately 30 percent less expensive than mail collection, while the monthly unit costs for CATI collection are 20 percent more expensive than mail (Clayton and Harrell 1990). Collection by TDE is now expanding in the CES, beginning with the establishments with the greatest employment. Research continues in determining ways to improve the touchtone system, even though respondents' acceptance of touchtone collection is highly favorable.

The purpose of the current research is to identify respondents' problems using the touchtone system and to measure errors associated with the mode. Results indicate where improvements are needed to reduce respondent problems as well as errors. Section 2 provides background information on TDE and the operations of the BLS TDE system. In the third section we address the potential errors due to the TDE mode of data collection, which requires a human-machine interface. In section 4 we describe the three sources of data collected for the study. A record-check survey, machine-recorded data and a respondent debriefing survey are used in the analysis of problems and errors. Section 5 includes the methods, analyses and findings from each of the data sources. Finally, in section 6 we provide an overall assessment of measurement errors due to this mode of collection, suggestions for improving the system, and implications for other surveys.

## 2. TOUCHTONE DATA ENTRY

The primary reason to consider collection by touchtone self-response is to reduce the cost of collecting data by CATI, while maintaining the timeliness and quality of CATI collection. The CES survey seemed to be a good candidate for touchtone collection, since only five or six numeric data items are collected each month. The data items include: all employees, women workers, production workers, production-worker payroll, production-worker hours, and for some industries, over-time hours or employee commissions. Establishments are asked to report totals for each data item for the pay period which includes the 12th of the month. TDE respondents are also asked to report their establishment identification number and the month for which they are providing data.

The features of the BLS touchtone system include the ability to:

- detect legitimate establishment respondents based on a match to a file of establishment numbers;
- vary the set of questions depending on the industry of the establishment;
- read back all responses for respondent confirmation using a computer simulated (digitized) voice (respondents are requested to enter “1” to confirm their answer or “0” to reenter their answer);
- wait two seconds for respondents to begin entering their answer, and wait two seconds between digits before interpreting data entry as complete (data entry is also assumed complete if the entire field length is filled – example: the data item “month” has a two-digit field);
- repeat each data item question up to three times (for identification number, month and all employees) or request the respondent to confirm that they have no answer for the question (for all other data items), if a respondent does not confirm their answer or if no answer is provided in the two seconds after the question is read;
- store the date, start and end time of each call, and all data items (Werking *et al.* 1988).



Respondents are mailed instructions on touchtone data entry. The instructions give direction on how to use the system and examples of the computer and respondent interaction, such as:

**Computer:**

Enter all employees  
You entered 2, 5

**Your Response:**

For 25 employees, press 2 and 5  
Press 1 to confirm, 0 to reenter.

The instructions also include the optional use of the “#” sign to indicate completion of data entry for an item. The use of the “#” sign reduces the time of the interview. In addition, before reporting their data respondents can call in to try out the system using a special test identification number.

Touchtone respondents are contacted by a telephone interviewer during their first month on the system to determine any problems they may have and to provide guidance, if necessary. Respondents receive a postcard reminder each month and a prompt call if they do not self-report by a specified date. The prompt call asks the respondent to telephone into the touchtone system as soon as possible. Data are not usually collected during the prompt call.

### 3. MEASUREMENT ERROR IN A HUMAN-MACHINE INTERFACE

Respondent use of a system such as TDE to answer survey questions has little precedent. However, touchtone recognition is widely used in such procedures as electronic banking and customer-controlled telephone services. While these services may save time and expense, they have a potential to alienate users. Problems and errors can originate with the system, task, or respondent. System problems primarily generate nonresponse error, while measurement error is related to the task and respondent performance.

While not directly related to surveys, the human-factors literature suggests several inter-related factors that may contribute to performance errors in a human-machine interface. First, respondents may not be familiar or comfortable with the technology. Waterworth (1984) suggests that the language of the human-machine interface is different than human communication as actions are performed in an order reflecting computer program logic. Since the ability to think in a way that parallels the logic is not a minor exercise, those with limited experience may have difficulty understanding the task and using the system. Second, synthetic speech is more difficult to understand than natural speech and places greater processing demands on working memory (Schwab, Nusbaum and Pisoni 1985). Thus, comprehension and memory problems associated with the mode may cause errors.

Synthetic speech includes both digitized speech, where a human voice is sampled, digitally encoded, and stored, and rule-based synthesized speech, generated using text as input (Marics and Williges 1988). The TDE system utilizes digitized speech, which is less difficult to understand than rule-based synthetic speech. However, comprehension problems occur with digitizing, as it introduces distortion into original speech (Cox and Coope 1981). Research shows that the understanding of synthetic speech may improve with training. In an experiment on perception of synthetic speech, Schwab and colleagues (1985) found that training with synthetic speech increases perception performance. Thus, comprehension may improve with exposure to and experience with the system. Another factor that may affect comprehension is the pace of the system. Marics and Williges (1988) found that the rate of the synthesized speech significantly affects speech intelligibility, as measured by transcription errors and response latency. However, subjects who received contextual information prior to listening to the speech had fewer transcription errors.

Thus, potential errors in the human-machine interface can occur from lack of experience with the technology and task, and from comprehension and memory problems associated with voice clarity and pace. Yet these problems are surmountable, as the evidence indicates that experience and training can improve performance.

#### 4. DATA

There were several objectives we considered in measuring TDE problems and mode error, and determining what data to use or collect. First, it was necessary to identify if and where problems were occurring. Second, we felt respondents should identify and interpret problems, but we also wanted measures independent of respondent assessment. Third, we needed to address problems and errors associated with the task and comprehension, including the possible improvement of respondent performance over time.

We decided to assess TDE problems and mode error using three different data sources, which have in common approximately 465 Pennsylvania business establishments. These establishments reported their monthly survey data by TDE to the Automated Collection Techniques (ACT) Laboratory at the BLS national office in Washington, D.C. A small number of the establishments began reporting by TDE to the ACT Lab in April, 1989. Others were added monthly through November of 1989. Most of the establishments continued reporting to the ACT Lab through April of 1990. The majority of these establishments moved from mail to TDE reporting.

The first source of data has two components. One is the TDE data recorded by machine from April to December, 1989. The other component is the same data recorded by establishments on a survey form. All respondents receive a yearly survey form on which they are requested to record their data for each month. Mail respondents fill in the form each month and mail it to the state employment security agency. The agency records the data, then returns the form by mail for next month's collection. CATI and TDE respondents are sent the survey form, but they do not return it. However, we sent a request to the TDE respondents to return their 1989 survey form, and obtained a 96 percent return rate. We then compared the TDE and form data, identifying discrepancies between the two. The TDE and survey-form data includes 1,930 observations across a nine-month period. Since establishments were phased into TDE slowly, the number of observations per establishment varies. The data cover approximately 75 establishments for 6-9 months, 200 establishments for 4-6 months and 190 establishments for 2-3 months. We refer to these data as the record-check data.

The second source of data includes machine-recorded information on respondent performance during the TDE telephone call. The TDE instrument was reprogrammed in January 1990 to automatically count and record the number of times a question was repeated due to nonresponse (question repeat), the number of times a respondent reentered data (data reentry) for each question, and the number of times an establishment called and hung up before entering data. Unfortunately, only the questions asking for the month and all employees total could be explicitly separated into question repeat and data reentry. For the data items including women and production workers, payroll and hours, we had to combine repeats and reentries, due to the structure of the original computer program. We refer to these data as the machine-recorded data.

The third source of data is a telephone debriefing survey, conducted from January to April of 1990 with the Pennsylvania establishments on their experiences with the TDE system. Approximately 411 business establishment respondents completed the interview, an 88 percent response rate. The questions covered such topics as voice quality, pace of interview, task problems, use of systems features, adequacy of instructional materials, and a system rating.

5. RESULTS

5.1 Record-Check Data

When we requested TDE establishments to return their survey forms, our first question was: how many respondents really used the forms? We speculated that one source of mode error was respondents who did not complete the form for use when entering their TDE data, which would increase demands on their memory. Those who did not complete the form might be more likely to enter and/or verify incorrect data. Thus, the request for the survey forms indicated that respondents were to return the form regardless of whether they completed it or not. However, of the 96 percent that returned their survey forms, only one establishment mailed in a blank form; all others sent in completed forms. While nonrespondents may work from memory, most of the respondents had completed their forms, giving us reason to believe memory problems due to lack of form use were not a major source of errors.

When comparing the data received by TDE with that on survey forms, we identified and coded discrepancies. The data on the survey form are those we would have received and used if the respondents were reporting by mail. The results are shown in Table 1. The first type of discrepancy occurred when the TDE data indicated there was no response for a data item, but there was a response on the establishment's survey form. This item nonresponse accounted for the greatest number of discrepancies, 82 out of 177, and was quite evenly spread across the applicable data items.

There was a pattern to the item nonresponse by month and establishment. The item nonresponse rate was 40 percent higher in the first month an establishment reported by TDE. In addition, some establishments had more difficulty than others, indicated by two or more nonresponses. Nearly half of all item nonresponse occurred in 18 establishments at or close to the time they began responding by TDE. This indicated problems existed with first-time use of TDE that might decrease with experience. Since the problem was concentrated in a small group, we believe it reflected lack of familiarity with automated processes. The remaining item nonresponse had no identifiable patterns; our suspicion was that some establishments simply missed the item, possibly due to office distractions, and continued on with the next question.

Table 1  
Record-Check Data - Number and Type  
of TDE Discrepancies

TDE item nonresponse	82
1-2 few/too many digits	18
Slipped on keypad	17
Dis/confirm - "1", "0" error	14
Form corrected, not TDE	12
No apparent error reason	26
Other reasons	8
Total	177



The second type of discrepancy was entering extra digits or, in a few cases, entering too few digits, which accounted for 18 of the 177 errors. This was specifically a problem associated with entering the payroll data item, where four respondents tried to enter cents instead of rounding to the nearest dollar. Several of the same respondents appeared to enter a half hour, 50, for production-worker hours rather than rounding. In the third type of discrepancy, the TDE numbers were nearly the same as those on the form, but one number off. The number entered incorrectly indicated a potential task problem in that the respondent may have had their fingers slide over on the keypad to the number directly on the side or below the correct digit. This accounted for 17 discrepancies. The fourth type of discrepancy occurred primarily for the all employee data item. There were eight establishments who had a "1" entered for this item in the TDE data, but had a larger employment number on their survey form. We speculate that respondents entered "1" twice when confirming the previous question on month.

Finally, there were a few respondents who had corrected data on their survey form, but not on TDE. There were other discrepancies which we could not explain. In addition, several respondents transposed their numbers or were off one category, accounting for the "other" reasons. For most of the errors, it was difficult to specifically ascertain if they were caused at the time of data entry, or not clearly comprehending the question or numbers being read back for verification. We suspected the former, but only for the second discrepancy, adding too many digits, could we really rule out comprehension problems.

The error rates for the survey items, ranging from 1.2 to 2.5 percent, are shown in Table 2. The all-employee, women-worker and production-worker questions have a lower percentage of errors than payroll and hours. This is not surprising since payroll and hours worked are usually four to six digits, compared to two to three digits for the other items. Thus, longer strings of numbers cause more difficulty. This may be related to difficulties entering the data, lack of respondent motivation in correction, or problems remembering longer strings of numbers during validation.

Table 3 shows the potential effect of the discrepancies on the CES data items, calculated by taking the sum of the difference between the values in the TDE system and the form, then dividing by the sum of the values on the form. The CES Survey uses a link-relative estimator for published estimates. The estimates in Table 3 do not take into consideration this estimator. However, the estimates in Table 3 provide an indirect measure of TDE mode error on survey estimates. None of the error is significantly different from zero at the five percent level. However, the potential for mode error appears to be more serious for production workers, payroll and production-workers hours. In this study, the number of production workers are overestimated by 7.3 percent, payroll by 7.3 percent, and hours by 4.4 percent.

Nearly all the discrepancies would have failed the edit parameters used in the CES survey and been corrected. The resultant effect of the discrepancies after edit corrections is zero for

Table 2  
Record-Check Data – Number of Discrepancies and Percent Error by Data Item

	All Employees	Women Workers	Production Workers	Payroll	Production Hours	Total
Discrepancies	23	29	28	48	49	177
% Error	1.2	1.5	1.5	2.5	2.5	1.8
(SE)	(.2)	(.3)	(.3)	(.4)	(.4)	(.3)

(N = 1,930 for each item).

Table 3

Record-Check Data – TDE Mode Error for Data Items Before and After Edit Corrections

	All Employees	Women Workers	Production Workers	Payroll	Production Hours
% Mode error, before edit corrections	0.0	0.5	7.3	7.3	4.4
(SE)	(.4)	(.3)	(5.2)	(3.8)	(3.7)
% Mode error, after edit corrections	0.0	0.0	0.0	0.0	0.0
(SE)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)

(N = 1,886–1,930 for each item).

all data items, as shown in Table 3. Examples of the large discrepancies include a respondent who incorrectly entered payroll for the number of production-workers, increasing production workers by over ten thousand, and two respondents who incorrectly entered the number of production-worker hours for number of production workers, raising the latter by several thousand. Payroll and hours have similar gross discrepancies, including the respondents who put in cents, instead of rounding for payroll.

5.2 Machine-Recorded Data

The touchtone system provides a tool for assessing difficulties which respondents have with this mode of collection. TDE can record the number of times respondents reenter their data, and how often the question is read back a second or third time before they respond. Information can also be kept on those respondents who hang-up before entering data.

The machine-recorded data were collected for a total sample of 1,203 observations over a three-month period in 1990. There were approximately 474 unique respondents, many of whom provided data for two or three months. There were few differences in the machine-recorded data by month, so all data are presented for the three months combined.

Figure 1 provides, for each data item, the percent of calls for which the question was stated to the respondent more than once. The question could be stated a second time if the respondent does not answer in two seconds (repeat), or if the respondent fails to confirm his or her answer by entering “1,” after it is read back (reenter). The figure indicates that the first two questions on the month and all employees, and the payroll and hours questions have higher rates of repeating and reentries by respondents than other data items. The higher rates for the first two questions – each over ten percent – may be due to respondents needing a few questions to orient themselves to the system. The payroll and hours questions generally have the greatest number of digits, so we suspect that data entry errors are more likely to occur, causing the question to be reread and the answer to be reentered.

Figure 2 provides data for the first two questions on month and employment for repeated questions after no answer (repeat) and after lack of confirmation (reenter), separately. It was not possible to acquire data separately for the other data items. The CES touchtone system requires respondents to enter at least their report identification number, the month and employment. The system will accept item nonresponse for the other data items. The mandatory entering of month and employment allowed the separation of repeated questions after no answer versus after a respondents lack of confirmation of the answer provided. Of respondents with problems



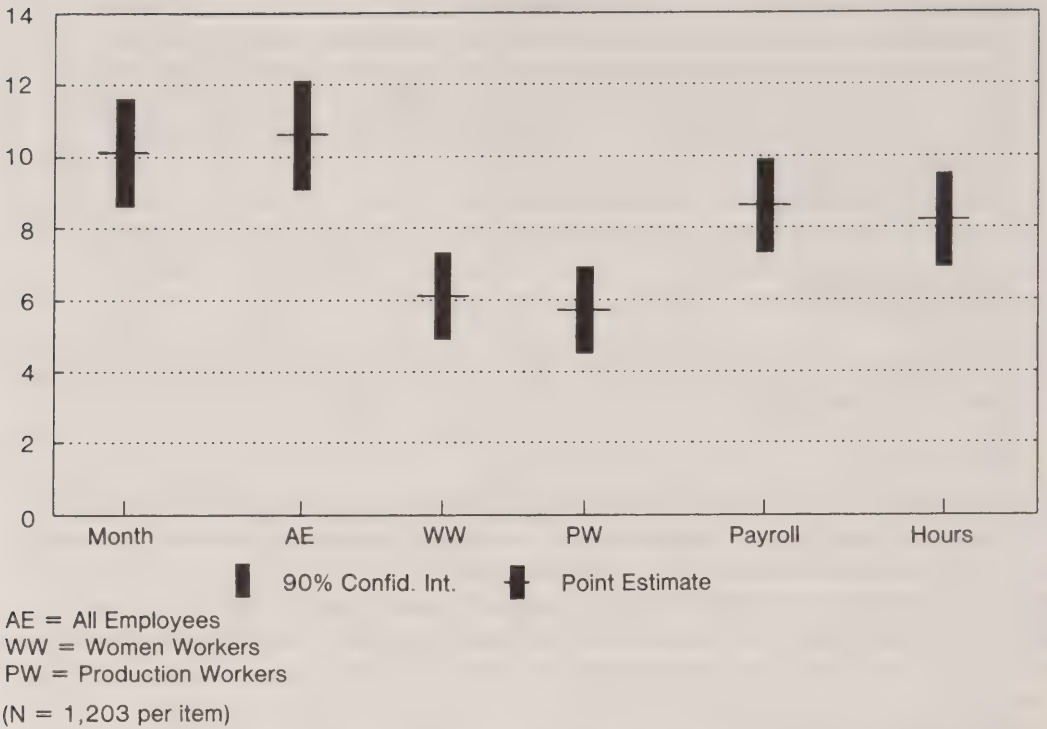


Figure 1. Machine Recorded Data - Percent of Questions Repeated/Reentered

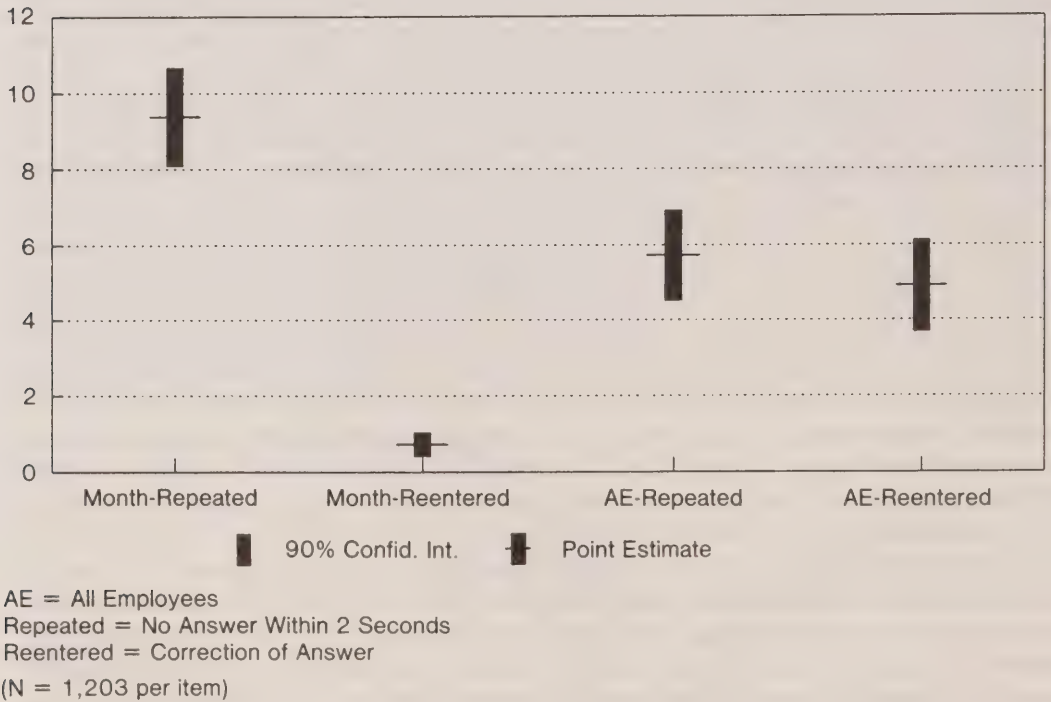


Figure 2. Machine Recorded Data - Percent of Questions Repeated Versus Reentered for 1st Two Data Items

on the month question, almost all were due to repeats, that is, two seconds passed without a response. On the other hand, problems with the employment question were almost evenly split between repeats after no previous answers and reentries after lack of confirmation of the previous answer.

Only two percent of the calls received by TDE each month included just a report identification number. During these calls, respondents had simply hung up, or could have been cut off the system.

In addition, the TDE system records all calls received that include at least the report identification number, month and employment. Using the TDE component of the record-check data discussed earlier, we identified respondents with more than one call during a month and coded reasons for the call-backs. In all, about four percent of the respondents called the system more than once in a given month. Most of these respondents provided data items which were not supplied on the initial call (2%). An additional one percent provided corrections to some data items in addition to new data items. Many of these respondents appeared to have had problems with entering the data the first time. Another one percent of the respondents called back only to provide corrections to data items previously supplied or provided identical data. These calls were often several days later, possibly implying that new data had been obtained from their records. In the case of the identical data, respondents may have forgotten whether or not they had previously reported their data. The system currently accepts the data with the latest date and time, although analysts are provided a list of respondents with duplicate records for review, and if necessary, correction.

A common reason for callback seemed to be related to the "enter 1" to confirm after each data item is entered. Many of the respondents who corrected their data had "1" in the data field prior to the callback, and some other response afterwards. Callbacks were twice as common with first-time respondents on the touchtone system than for respondents who were "experienced" users.

A few respondents called in their data three times for a given month, and one respondent called in data four times. These respondents seemed to be having difficulties with the system, but finally reported all of their data correctly.

Overall, these data suggest that respondents are having some difficulties with the system (more than they admit to during the respondent debriefing interview). Some steps could be taken to help alleviate some of the problems. These include providing more time for respondents to answer, providing better instructions, and trying to improve the confirmation of data entry method. In addition, being able to go back to a data item might solve some of the problems.

### 5.3 Respondent Debriefing Survey

BLS interviewers conducted a telephone debriefing survey with TDE respondents during 1990. Given the human-factors literature discussed earlier, some of the questions focused on understanding and pace of the digitized voice. The results from the machine-recorded data showed a substantial number of repeats and retries of questions, thus, questions were developed to address that topic. In addition, respondents were asked to rate the TDE system and answer questions relating to systems design.

The results from the debriefing survey are presented in Table 4. Respondents expressed little difficulty in comprehending the digitized voice. About 97 percent said the voice was very understandable, and all respondents indicated that it was easy to understand the numbers as the voice read them back for confirmation. During the first two months of the survey, we asked respondents about the pace of the interview. Most of the respondents said the pace was about right (88%), although about ten percent felt it was too slow. Our suspicion throughout the

**Table 4**  
Results of Debriefing Survey\*

Voice understandable	97%
Easy to understand #'s read back	100%
Pace about right	88%
Never reentered numbers	60%
Never repeated questions	83%
Never had poor telephone connection	93%
Used speed enhancement feature	63%
Instructions adequate	98%
TDE experience very favorable	93%

\* N = 411, except for the pace and question repeat items. Approximately 177 respondents were asked about pace and 209 were asked about the repeating of questions.

study was that voice comprehension was much less a problem than were difficulties carrying out the task. While it was difficult to separate out the two in the record-check data, the debriefing interviews lend support to our suspicion.

For task difficulties, 60 percent of respondents said they never had to reenter numbers, while most of the others indicated they had to reenter numbers sometimes. When asked the reasons for reentering numbers, a majority indicated they had accidentally entered a wrong number. Others said they did not have enough time, were distracted, or entered their numbers too fast. In the latter several months of interviewing, respondents were asked about the repeating of questions (without reentering data). About 83 percent of the respondents said they never found it necessary to repeat questions. Of the 17 percent who repeated questions, the majority said they were distracted, while others said they did not have enough time.

Most respondents had little difficulty with telecommunications failure, as 93 percent said they never experienced a poor telephone connection when using TDE. Of the respondents who did get a poor connection, most said it happened only once. A large number of the respondents, 63 percent, used the pound sign, a feature of the system designed for speeding up the reporting of data.

Nearly all respondents said the instructions sent to them as they began TDE were adequate. Overall, respondents seemed satisfied with the TDE system – approximately 93 percent rated their experience using TDE as very favorable.

**6. DISCUSSION**

The data show few serious problems with the TDE mode of data collection. Record-check data indicate some item nonresponse error, which is associated with first-time users. Entering additional or incorrect digits appears to be the most serious problem affecting the data items. However, in a panel survey, longitudinal edit checks could reduce this error, as could logical edit checks in all surveys. In addition, the rounding of data needs to be addressed in respondent instructions. Both the record-check and machine-recorded data show that there are more difficulties with longer strings of numbers, probably in both entering data and verifying



incorrect data. The latter could indicate difficulty remembering longer number chains during verification, as comprehension of numbers appeared to be good, *i.e.* respondents said they easily understood numbers being read back for confirmation.

Record-check data show that establishments may have carried over their confirmation of the month into the all-employee question. In addition, the machine-recorded data indicate respondents often do not respond to the month question the first time it is asked. Since respondents appear to be using their survey forms as they enter data, it is likely that moving from the identification number at the top of the form, to the month and data items further down the form, they require extra time to locate themselves. This problem could be solved by placing all information that needs to be entered in one location on the survey form. This might reduce the number of question repeats for the "month" item and potentially lower costs by reducing the length of calls. Question repeats for other items might be reduced by giving respondents more time to respond, since they report they were distracted from the task. However, since most respondents feel the pace of the system is about right, and many are using the speed enhancement feature, adding more time could cause frustration. Probably little can be done to reduce the number of reentries, as respondents indicate they have entered a wrong number and need to correct it.

The data show that errors are reduced with experience. This indicates that a panel survey may be best for this mode of data collection. For surveys requiring numeric or yes/no responses, we believe touchtone also has great potential. The errors are not extremely serious, and respondents rate their experiences using TDE very favorably. TDE may be particularly attractive to business respondents, who can call at convenient times, rather than be interrupted by telephone calls requesting data. However, for some surveys, self initiation and the lack of human contact may be problems which would contribute to nonresponse error.

Although respondent acceptance of touchtone collection is very favorable, there are some steps which can be taken to make the system better. These include:

- giving respondents enough time to key enter their data, especially for the first few questions and those which have a long string of digits,
- investigating ways to improve the confirmation of data items, and
- providing longitudinal edit checks to detect reporting of dollars and cents and other gross errors. The edits could be built into the TDE system with appropriate questions/probes to respondents to correct or confirm their answers.

BLS has used touchtone collection with one other survey. This survey was a small sample follow-up of business establishments who had participated in a Survey of Employer Drug Assistance Programs in 1988. The follow-up survey in 1990 was intended to determine if any substantial changes had occurred in the percentage of establishments providing employer drug assistance programs over the past two years. These establishments were mailed a short survey questionnaire requesting numeric or yes/no answers and encouraged to report their data by touchtone telephone. At the end of the first several weeks of the survey, approximately 20 percent of the establishments had reported their data by touchtone, and an equal amount by mail. TDE was not used after nonresponse follow-up activities began - about two weeks after the initial mailout. The remaining data were collected by telephone (CATI).

We believe that other surveys with time dependent data can take advantage of the time and keypunch savings of touchtone data collection. The mode may communicate the importance of timeliness to the respondent. This paper indicates that measurement errors are controllable using touchtone collection.

Given the timeliness and lower costs of touchtone data collection, we expect it will be used more extensively in the future. We know of two current projects testing touchtone recognition in a survey setting. Statistics Canada is testing a touchtone data collection system for the Survey of Employment, Payroll and Hours. In addition, a touchtone system for a survey of AT & T customers is being developed at Bell Laboratories (Wendler 1990).

BLS is also experimenting with the use of voice recognition technology for data collection in the CES survey (see Winter and Clayton 1990). While touchtone telephones are increasingly available, we estimate that between 10 to 20 percent of our respondents do not have touchtone telephones. Once speaker-independent voice recognition technology reaches an acceptable level for the ten digits needed to report CES data, we expect users will prefer it over touchtone collection. Further work on measurement errors associated with voice recognition technology needs to be undertaken.

### ACKNOWLEDGEMENTS

The authors thank Darrell Philpot and Henry Chiang for their assistance with this research.

### REFERENCES

- CLAYTON, R., and HARRELL, L.J., Jr. (1990). Developing a cost model for alternative data collection methods: Mail, CATI and TDE. Presented at the Annual Meeting of the American Statistical Association, Anaheim, California.
- COX, A.C., and COOPE, M.B. (1981). Selecting a voice for a specified task: The example of telephone announcements. *Language and Speech*, 24, 233-243.
- MARICS, M.A., and WILLIGES, B.H. (1988). The intelligibility of synthesized speech in data inquiry systems. *Human Factors*, 30, 719-732.
- PONIKOWSKI, C.H., COPELAND, K.R., and MEILY, S.A. (1989). Applications for touch-tone recognition technology in establishment surveys. Presented at the American Statistical Association Winter Conference, San Diego, California.
- SCHWAB, E.C., NUSBAUM, H.C., and PISONI, D.B. (1985). Some effects of training on the perception of synthetic speech. *Human Factors*, 27, 395-408.
- WATERWORTH, J.A. (1984). Interaction with machines by voice: A telecommunications perspective. *Behaviour and Information Technology*, 3, 163-177.
- WENDLER, E.R. (1990). Respondent-initiated computer-directed surveys. Presented at the Annual Conference of the American Association of Public Opinion Research, Lancaster, Pennsylvania.
- WERKING, G.S., TUPEK, A.R., PONIKOWSKI, C., and ROSEN, R. (1986). A CATI feasibility study for a monthly establishment survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 639-643.
- WERKING, G., TUPEK, A., and CLAYTON, R. (1988). CATI and touchtone self-response applications for establishment surveys. *Journal of Official Statistics*, 4, 349-362.
- WINTER, D.L.S., and CLAYTON, R.L. (1990). Speech data entry: Results of the first test of voice recognition for data collection. Bureau of Labor Statistics, Washington, D.C.

# Avoiding Sequential Sampling with Random Digit Dialing

J. MICHAEL BRICK and JOSEPH WAKSBERG<sup>1</sup>

## ABSTRACT

The Mitofsky-Waksberg procedure is an efficient method for selecting a self-weighting, random digit dialing (RDD) sample of households. The Mitofsky-Waksberg procedure is sequential, requiring a constant number of households be selected from each cluster. In this article, a modified Mitofsky-Waksberg procedure which is not self-weighting or sequential is described. The bias and variance for estimates derived from the modified procedure are investigated. Suggestions on circumstances which might favor the modified procedure over the standard Mitofsky-Waksberg procedure are provided.

**KEY WORDS:** Random digit dialing; Telephone sampling; Cluster sampling; Trimming.

## 1. INTRODUCTION

The Mitofsky-Waksberg procedure for selecting random digit dialing samples of households (Waksberg 1978) is frequently used for sample selection in telephone surveys. As described in the Waksberg paper, it is an efficient method of producing a self-weighting sample, that is, one in which all telephone households have the same probability of selection (except for households with more than one telephone number). The efficiency is due to the sharp reduction in the proportion of nonhousehold telephone numbers that have to be dialed in order to identify sample households.

The Mitofsky-Waksberg procedure is a two-stage sample design. In the first stage, a sample of clusters is chosen where the clusters consist of blocks of 100 telephone numbers, or multiples of such blocks. The clusters (or blocks of 100 telephone numbers) are first selected with equal probability. One telephone number is chosen at random in each cluster and dialed. If the number is that of a household, the cluster is retained. Otherwise, it is rejected. The second stage is the selection of households within the retained sample clusters. For the self-weighting feature of the sample to apply, a constant number of households per cluster is required. Some organizations (including Westat Inc.) generally go a little further and specify a constant number of interviewed households per cluster (or screened households if the first part of data collection is screening). The rationale is that substituting another randomly selected household within the same cluster for each nonrespondent is a reasonable way of reducing nonresponse bias.

There is an awkward operational feature to this system. It sometimes takes a fairly large number of callbacks to determine whether or not a telephone number is residential, particularly for numbers that repeatedly ring with no answer. Even more are needed to learn which households cooperate. Such determinations must be made for an initially selected sample to ascertain which clusters require more telephone numbers to achieve the desired cluster size and how many telephone numbers have to be added. In effect, a sequential scheme is necessary for each cluster, where all previous cases need to be cleared up before it is known whether the sample needs to be increased. This process is particularly inconvenient when there is a tight time schedule for data collection.

Several attempts to modify the Mitofsky-Waksberg method have been proposed which reduce or eliminate the sequential features of the plan. Potthoff (1987) developed a generalization

<sup>1</sup> J. Michael Brick and Joseph Waksberg, Westat Inc., 1650 Research Boulevard, Rockville, Maryland 20850, U.S.A.



of the Mitofsky-Waksberg technique in which  $c$  telephone numbers are chosen per cluster in determining whether to retain the cluster, whereas Mitofsky-Waksberg use only one. A self-weighting sample is achieved by having clusters in which only one of the  $c$  telephone numbers dialed continue with a sampling plan that includes having a fixed number of households per cluster, and the remaining clusters having a fixed number of telephone numbers. The latter group of clusters does not require a sequential approach. Potthoff reports that in practice most clusters will fall into the second class so that the sequential operations, although not eliminated, are sharply reduced.

Lepkowski and Groves (1986) describe a sampling method in which blocks of telephone numbers with more than a trivial number of telephone numbers listed in directories (and other sources, if available) are selected with probability proportionate to the number of listed numbers. Blocks of numbers which contain zero or very few telephone numbers are sampled through the Mitofsky-Waksberg procedure. Sudman (1973) had previously proposed sampling blocks of numbers with probability proportionate to the numbers listed in directories, but without making any provision for empty blocks (which could have unlisted numbers). Drew and Jaworski (1986) describe an RDD survey carried out in Canada in which purchased counts of residential numbers (both published in directories and nonpublished) were used as measures of size. Since the counts were considered as virtually complete, there was no need to sample empty blocks. As far as we are aware, there is no way of getting virtually complete counts of residential numbers in the U.S.

Neither the Potthoff nor the Lepkowski-Groves sample design completely eliminates the need for a sequential process, although both appear to reduce the portion of the sample which requires it. There are also some other disadvantages to the two procedures. The Potthoff technique appears to be rather complex – as far as we know it has not been used much for RDD surveys. For national surveys, the Lepkowski-Groves technique involves the purchase of a directory list covering the total U.S. and processing it to obtain measures of size. Such commercial lists are available, but they are expensive. Furthermore, a number of recent reports indicate the percentage of all residential numbers that are listed in directories is not very high, and is rapidly decreasing. Tucker (1989) describes an analysis of listed numbers in a group of U.S. counties and cities which shows listing rates varying from 48 to 62 percent. An article by Linda Piekarski (1990) states that if the rate of increase of unlisted numbers continues at the current level, “as many as 62% of the nation’s households may be unlisted by the year 2000.” The measures of size thus are probably only moderately correlated with the actual number of households in a working block.

Waksberg has suggested an alternative modification of the Mitofsky-Waksberg procedure (Waksberg 1984) which completely eliminates the need for sequential sampling. Westat has used this method in a large number of studies using RDD. Cummings (1979) had previously used the same procedures as a result of an error in implementing the Mitofsky-Waksberg procedure. Cummings did not recognize its usefulness in avoiding sequential sampling and did not explore its features for use in other surveys. We describe the method and its mathematical and statistical properties.

## **2. ALTERNATIVE METHOD OF ESTABLISHING CLUSTER SIZES WITH MITOFSKY-WAKSBERG TECHNIQUE**

As indicated earlier, the Mitofsky-Waksberg technique requires a constant number of sample residential numbers per cluster (or block of numbers) to produce a self-weighting sample. The alternative that is proposed is to use a constant number of telephone numbers per cluster for

the sample ( $K$ ). The first stage of selection is unchanged. (The first-stage selects clusters with probability proportionate to the number of households.) With a constant number of telephone numbers per cluster, the sample numbers can be designated in advance eliminating the sequential process. We note that followup effort is still necessary to determine which sample telephone numbers are residential, both in the first and in the second stages of sampling. However, this has to be done for a fixed set of telephone numbers. A sequential process is not involved.

The alternative procedure does not produce a self-weighting sample. Since the first stage is selected with PPS, the probability of a cluster being selected is  $r N_i/100$  where  $r$  is the sampling rate for selection of the clusters, that is, the first stage selection rate, and  $N_i$  is the number of residential numbers in the  $i$ th cluster. The weight should be proportional to  $N_i^{-1}$ , but since  $N_i$  is not known, it is taken to be proportional to  $n_i^{-1}$ , the number of sample households in the cluster.

This modification of the Mitofsky-Waksberg method has good features for survey operations. It is simple. The sample can be virtually preselected and no costly control operations are needed. Although weighting is required, the weights are directly available from the sample data, and they can be mechanically produced without any extensive professional oversight.

There are, however, some serious problems. First, there is a bias when  $N_i^{-1}$  is estimated by  $K/100n_i$  where  $K$  is the number of telephone numbers selected per cluster (a constant number in all clusters). The bias is fairly small, but it does exist. It cannot be eliminated or reduced by minor modifications of the weights, such as using  $1/(n_i + t)$  instead of  $n_i^{-1}$ , with “ $t$ ” denoting a fixed constant. Secondly, the introduction of variable weights increases the variances of the estimates substantially. (The increase is not so much caused by the weights as by the fact they reflect variable probabilities of selection.) Finally, the modification loses one of the useful features of the Mitofsky-Waksberg method – the ability to fix the exact sample size desired. The Mitofsky-Waksberg method’s use of a constant number of households per cluster means that any desired sample size can be obtained by selecting a sample with the appropriate number of clusters. With the modification, the sample size becomes a random variable, which generally will not be exactly equal to the desired sample size. Although the deviations are usually small, the ability to achieve exact target sizes is useful when contracts or budget commitments require the survey organization to satisfy exact target requirements. We discuss these issues in Sections 3 and 4.

Before going on to a discussion of the variances and biases, it is useful to examine the distribution of cluster sizes in the U.S. Tables 1 to 3 show estimates of such distributions prepared from data reported in two large national U.S. surveys conducted via RDD by Westat Inc. Both of these surveys used the modification of the Mitofsky-Waksberg procedure described above. The sample for the survey summarized in Table 1 was selected in 1986 and consisted of 2,427 clusters (retained after first-stage sampling) with 15 telephone numbers per cluster, or 36,405 total numbers. There were 18,756 completed screeners, 2,396 refusals, 1,727 nonresponse for other reasons, and 13,526 nonresidential or nonworking numbers, ring no answers, and cases that could not be classified. The analysis is restricted to the 18,756 completed cases. The data in Tables 2 and 3 are based on a 1989 sample of 1,000 clusters with 30 telephone numbers per cluster or 30,000 telephone numbers, of which 19,586 were residential with screeners completed. Table 2 shows the distribution of the 15,030 completed cases and Table 3 shows the distribution of the 19,586 residential numbers found in the 1,000 clusters. The cluster weights shown are expressed as  $\bar{n}/n_i$  where  $\bar{n}$  is the average number of households per cluster. It seems useful to express them in this form since they then show the deviations from a self-weighting sample. The design effects only account for the increased variances arising from variable sampling fractions. They do not include effects of other aspects of the sample design.



**Table 1**  
Number of Completed Screeners per Cluster in 1986 Survey  
(Based on sample of 2,427 clusters with 15 telephone numbers per cluster)

Number of Completes per Cluster	Average Cluster Weight <sup>1</sup>	Household Distribution			Cluster Distribution		
		Frequency	Percent	Cumulative Percent	Frequency	Percent	Cumulative Percent
0	xx	0	0	0.0	62	2.6	2.6
1	7.93 <sup>2</sup>	54	0	0.3	54	2.2	4.8
2	3.97 <sup>2</sup>	106	0.6	0.9	53	2.2	7.0
3	2.64	258	1.4	2.2	86	3.5	10.5
4	1.98	440	2.3	4.6	110	4.5	15.0
5	1.59	810	4.3	8.9	162	6.7	21.7
6	1.32	1,290	6.9	15.8	215	8.9	30.6
7	1.13	1,960	10.5	26.2	280	11.5	42.1
8	0.99	2,656	14.2	40.4	332	13.7	55.8
9	0.88	2,862	15.3	55.6	318	13.1	68.9
10	0.79	2,990	15.9	71.6	299	12.3	81.2
11	0.72	2,717	14.5	86.1	247	10.2	91.4
12	0.66	1,548	8.3	94.3	129	5.3	96.7
13	0.61	780	4.2	98.5	60	2.5	99.2
14	0.57	210	1	99.6	15	0.6	99.8
15	0.53	75	0	100.0	5	0.2	100.0
Total		18,756	100.0	xx	2,427	100.0	xx
Mean cluster size <sup>3</sup>			7.93				
Design effect <sup>4</sup>			1.31				

<sup>1</sup> The cluster weight is the mean cluster size (i.e., 7.93) divided by the number of completes in the *i*-th cluster.  
<sup>2</sup> Trimming the weights would bring these weights down to 3.  
<sup>3</sup> The mean cluster size is the average over the 2,365 clusters with one or more completed screeners.  
<sup>4</sup> The design effect is reduced to 1.12 if the maximum weight is 3.

It should be noted that Table 1 is based on a sample of 15 telephone numbers per cluster and Tables 2 and 3 used 30 telephone numbers per cluster. Estimates of the percent residential in a cluster based on 15 telephone numbers will, of course, be subject to a higher sampling error than an estimate based on 30 telephone numbers. However, the number of clusters used in Table 1 was more than twice those in Tables 2 and 3 which should largely offset the effect of the different cluster sizes.

There are two differences between Tables 2 and 3. One is that Table 2 shows the distribution of completed screeners (as does Table 1) while Table 3 is based on all sample households. The use of only completed cases in Table 2 reduces the estimate of the average number of households per cluster and shifts the entire distribution. In addition, it introduces more variability to the estimates of the distribution shown because the distributions reflect sampling errors of both the distribution of households per cluster and the distribution of nonresponse rates per cluster. The second difference is that Table 2 (and Table 1) is expressed in terms of the number of cases per cluster and Table 3 shows the distributions by the percentage of residential numbers per cluster. It was convenient to express Table 3 in that form for analyses described later in this report.



One other feature of the percentages shown in Tables 1 to 3 should be noted. They reflect the size distributions of clusters which fell into the sample, not the distribution of clusters in the U.S. The use of probability proportionate to size sampling results in an oversampling of clusters with a high proportion of residential numbers and an underrepresentation of clusters with a small number. It is possible to convert the distribution from one that represents the sample to one that represents the population by multiplying each percentage by the cluster weights and computing the percentage distribution of the resulting figures. Since the weights are exactly proportional to the reciprocal of the number of completes per cluster, it turns out that converting the household distribution so that it represents the distribution in the population produces the percentages shown in the cluster distribution. The cluster distribution in the sample is thus the same as the household distribution in the population.

We show distributions of both all-sample households and completed cases because both are of interest to researchers. The Table 3 data have been used for the analyses in Sections 3 and 4.

**Table 2**  
Number of Completed Screeners per Cluster in 1989 Survey  
(Based on sample of 1,000 clusters with 30 telephone numbers per cluster)

Number of Completes per Cluster	Average Cluster Weight <sup>1</sup>	Household Distribution			Cluster Distribution		
		Frequency	Percent	Cumulative Percent	Frequency	Percent	Cumulative Percent
0	xx	0	0	0.0	8	0.8	0.8
1 or 2	7.57 <sup>2</sup>	6	0	0.0	3	0.3	1.1
3 or 4	4.33 <sup>2</sup>	37	0.2	0.3	10	1.0	2.1
5 or 6	2.75	126	0.8	1.1	22	2.2	4.3
7 or 8	2.02	403	2.7	3.8	53	5.3	9.6
9 or 10	1.59	688	4.6	8.4	72	7.2	16.8
11 or 12	1.32	1,325	8.8	17.2	115	11.5	28.3
13 or 14	1.12	1,987	13.2	30.4	147	14.7	43.0
15 or 16	0.98	2,636	17.5	50.0	170	17.0	60.0
17 or 18	0.85	2,692	17.9	65.9	154	15.4	75.4
19 or 20	0.78	2,387	15.9	81.8	123	12.3	87.7
21 or 22	0.70	1,673	11.1	92.9	78	7.8	95.5
23 or 24	0.64	816	5.4	98.3	35	3.5	99.0
25 or 26	0.55	254	1.7	100.0	10	1.0	100.0
27 or 28	xx	0	0	100.0	0	0	100.0
29 or 30	xx	0	0	100.0	0	0	100.0
Total	xx	15,030	xx	xx	1,000	xx	xx
Mean cluster size <sup>3</sup>				15.11			
Design effect <sup>4</sup>				1.33			

<sup>1</sup> The cluster weight is the mean cluster size (*i.e.*, 15.15) divided by the number of completes in the *i*-th cluster.

<sup>2</sup> Trimming the weights would bring these weights down to 3.

<sup>3</sup> The mean cluster size is the average over the 992 clusters with one or more completed screeners.

<sup>4</sup> The design effect is reduced to 1.12 if the maximum weight is 3.

**Table 3**  
Proportion of Residential Numbers per Cluster in 1989 Survey  
(Based on sample of 1,000 clusters with 30 telephone numbers per cluster)

Proportion of Residential nos. per Cluster	Average Cluster Weight <sup>1</sup>	Household Distribution			Cluster Distribution		
		Frequency	Percent	Cumulative Percent	Frequency	Percent	Cumulative Percent
0	xx	0	0.0	0.0	5	0.5	0.5
.001 to .049	21.76 <sup>2</sup>	5	0.0	0.0	3	0.3	0.8
.05 to .099	8.70 <sup>2</sup>	18	0.1	0.1	6	0.6	1.4
.10 to .149	5.22 <sup>2</sup>	41	0.2	0.3	9	0.9	2.3
.15 to .199	3.73 <sup>2</sup>	48	0.2	0.6	8	0.8	3.1
.20 to .249	2.90	53	0.3	0.8	7	0.7	3.8
.25 to .299	2.37	144	0.7	1.6	16	1.6	5.4
.30 to .349	2.01	178	0.9	2.5	17	1.7	7.1
.35 to .399	1.74	408	2.1	4.6	34	3.4	10.5
.40 to .449	1.54	459	2.3	6.9	34	3.4	13.9
.45 to .499	1.37	840	4.3	11.2	56	5.6	19.5
.50 to .549	1.24	1,040	5.3	16.5	63	6.3	25.8
.55 to .599	1.14	1,926	9.8	26.3	107	10.7	36.5
.60 to .649	1.04	2,126	10.9	37.2	109	10.9	47.4
.65 to .699	0.97	3,255	16.6	53.8	155	15.5	62.9
.70 to .749	0.90	2,610	13.3	67.1	116	11.6	74.5
.75 to .799	0.84	3,022	15.4	82.6	126	12.6	87.1
.80 to .849	0.79	1,556	7.9	90.5	61	6.1	93.2
.85 to .899	0.75	1,458	7.4	98.0	54	5.4	98.6
.90 to .949	0.71	399	2.0	100.0	14	1.4	100.0
.95 to .999	xx	0	0.0	100.0	0	0.0	100.0
Total	xx	19,586	100.0	xx	1000	100.0	xx
Mean cluster size <sup>3</sup>			19.68				
Design effect <sup>4</sup>			1.28				

<sup>1</sup> The cluster weight is the mean proportion in a cluster (*i.e.*, 0.653) divided by the proportion of residential numbers in the *i*-th cluster.  
<sup>2</sup> Trimming the weights would bring these weights down to 3.  
<sup>3</sup> The mean cluster size is the average over the 995 clusters with one or more residential numbers.  
<sup>4</sup> The design effect is reduced to 1.12 if the maximum weight is 3.

**3. VARIANCE IMPLICATIONS OF THE MODIFIED  
MITOFSKY-WAKSBERG METHOD**

In the standard Mitofsky-Waksberg method the variance of a sample estimate is dependent upon the number of households selected per cluster and the homogeneity of the households within and between clusters. The variance for a cluster sample can be written as the variance for a simple random sample multiplied by  $[1 + \rho(\bar{n} - 1)]$ , where  $\rho$  is intraclass correlation and  $\bar{n}$  is the average number of households per cluster. Since telephone clusters are often related to geographic areas and tend to be somewhat homogeneous, selecting a large number of households per cluster can be inefficient.

When the modified Mitofsky-Waksberg method is used, another source of variance is introduced because the number of households selected per cluster is allowed to vary from cluster to cluster. As pointed out in Section 2, the denominator of the second stage probability of selection does not cancel with the number of households in the cluster (which is proportional to the probabilities in the first stage) and the overall probabilities of selecting households vary from cluster to cluster.

The variability among clusters in the overall household sampling rates causes the variances of the estimates to be larger than those in the standard Mitofsky-Waksberg method where each household has the same probability of selection. Methods for estimating the increase in the variance of an estimate arising from unequal probabilities of selection are discussed by Kish (1965) and by Waksberg (1973). A simple approximation to the variance of an estimate under an unequal weighting scheme (where the weights do not reflect variable sampling rates in strata deliberately chosen to reduce sampling variances) is the sampling variance which would occur with a self-weighting sample multiplied by a variance inflation factor (VIF), given by  $VIF = \{1 + Relvar(weights)\}$ . We will use this approximation below to investigate the variance implications associated with the modified Mitofsky-Waksberg method.

The relative variance of the weights was computed by partitioning the process into two components. First, the mean and variance of the weights were computed conditioned upon sampling from a truncated (since zero households cannot be obtained if the cluster is sampled in the first stage) hypergeometric distribution, defined by the household density in the cluster and the cluster sample size. The unconditional mean and variance of the weights were then computed by integrating over the distribution of households in the sampled clusters shown in Table 3. The distribution of households in the sample is critical in the evaluation of the VIF.

The natural weight assigned to a household in the modified Mitofsky-Waksberg is proportional to  $n_i^{-1}$ , where  $n_i$  is the number of households observed in sample cluster  $i$ . This weight can vary by factors which range from as little as  $1/K$  to 1, where  $K$  is the number of telephone numbers selected in a cluster. The average weight is roughly  $1.5/K$ , since about 65 percent of numbers in the sampled clusters are residential.

If the number of telephone numbers sampled per cluster is between 5 and 30, then the increase in variance due to the weighting is about 30 percent. The VIF decreases slightly as the number sampled per cluster increases beyond 30, reaching approximately 17 percent when all the numbers in the cluster are sampled.

The VIF or the relative variance of the weights is a function of the distribution of the number of households across clusters and random sampling variability within the clusters. This decomposition is made explicit by expressing the variance of the weights as the sum of the mean of the conditional variance of the weights and the variance of the conditional mean of the weights, where the conditioning is with respect to the household density of the cluster.

When the cluster sample size is small, the mean of the conditional variance is the dominant component of the overall variance. As the cluster sample size increases, the variance of the conditional mean becomes more dominant. This is why the relative variance of the weights, shown in the first row of Table 4, is not a monotonic function of the cluster sample size.

Table 4  
Approximate Variance Inflation Factors (VIF) for Modified  
Mitofsky-Waksberg Random Digit Dialing Samples

Weight	Cluster Sample Size ( $K$ )				
	5	10	30	60	100
$1/n_i$	1.31	1.34	1.29	1.23	1.17
$1/(n_i + .5)$	1.18	1.21	1.20	1.18	1.16
$1/(n_i + 1)$	1.12	1.15	1.16	1.15	1.14
$1/(n_i + 2)$	1.07	1.09	1.11	1.12	1.13



Variances Using Different Weights

Weights other than ones proportional to the inverse of the number of households were also examined to determine their impact on the bias and variance of the estimates. Many of the alternative weights studied were derived from variance stabilizing transformations suggested for binomial variables.

Of all the alternatives examined, the estimators with the best bias and variance properties involved simple adjustments of the natural weight. In particular, adding a small constant to the observed number of households (estimators of the form  $(n_i + t)^{-1}$  where  $t$  is .5, 1, or 2) resulted in reducing the increases in variance due to differential weighting. The addition of the constant reduces the range of the weights by cutting the values of the largest weights while only slightly modifying the weights for clusters where more households are found in the sample.

Table 4 shows the VIF for the estimators of the form  $(n_i + t)^{-1}$  for different numbers of telephone numbers sampled per cluster. The table also is based on the household and cluster distributions shown in Table 3. It is clear from the table that a substantial reduction in the variance due to unequal weighting can be achieved by using  $(n_i + 1)^{-1}$ , rather than the natural estimator. This is especially true for RDD designs which sample 30 or fewer telephone numbers per cluster. The increase in variance due to differential weighting for  $(n_i + 1)^{-1}$  is only 16 percent when 30 numbers are selected per cluster as opposed to a 29 percent increase when the natural estimator is used.

Variances with Trimmed Weights

A practice that is often used to mitigate the variance inflation associated with varying weights is the truncation of very large weights. This truncation, or trimming of weights, is usually fixed at a weight above which relatively few observations are found. In many Westat RDD samples, weights that exceed two or three times the mean weight have been truncated. For this research, we have examined weights truncated at about 3 times the mean weight. For samples of 10 per cluster, the weights were truncated at 2 times the mean weight because so few observations are affected otherwise.

Table 5 shows the VIF for the estimators for different cluster sample sizes when the weights are trimmed at three times the mean weight for  $n_i^{-1}$ . The VIF's for samples of 5 per cluster are not given because the truncation point in samples of this size is nearly at unity, the largest possible weight.

Table 5  
Approximate Variance Inflation Factors (VIF) for Modified Mitofsky-Waksberg  
Random Digit Dial Samples with Trimmed\* Weights

Weight	Cluster Sample Size ( <i>K</i> )			
	10	30	60	100
1/ <i>n<sub>i</sub></i>	1.12	1.11	1.09	1.09
1/( <i>n<sub>i</sub></i> + .5)	1.11	1.10	1.09	1.09
1/( <i>n<sub>i</sub></i> + 1)	1.09	1.10	1.09	1.09
1/( <i>n<sub>i</sub></i> + 2)	1.07	1.09	1.09	1.08

\* All weights trimmed at 3 times the mean weight, except samples of 10 trimmed at 2 times the mean.

The tabled values show trimming substantially reduces adverse impact of the differential weights on the variance of the estimates. The most dramatic reduction is for the natural estimator; its VIF is reduced by over 50 percent by the use of trimming. The VIFs for the other estimators are improved somewhat, but the reductions are less striking since they already had smaller VIF's than the natural estimator. Trimming has the potential of introducing biases which may counteract the advantage in variance reduction. Biases are discussed in Section 4.

### Variances with Augmented Sampling

A third approach to reducing the variability of the weights is the use of augmented sampling. Large weights occur when the number of households identified in the cluster is small relative to the expected number of households per cluster. To reduce the chance for this happening, an augmented sampling procedure can be used. If the number of households in a cluster is smaller than a fixed number (say less than one third of the mean number per cluster), then the sample size in the cluster can be doubled or increased by some other amount.

This procedure could be iterated to insure that the number of households per cluster reaches a specified limit or until all numbers in the cluster are used. The obvious disadvantage of this iterative plan is that it requires monitoring sample yield by cluster and the very fact that it is sequential. Another disadvantage of the method is that it results in sampling more telephone numbers from clusters that have a lower household density (the ones most likely to need augmentation), hence reducing productivity.

Despite the operational shortcomings of the augmented sampling approach, we did a limited examination of the method. Since the results for the augmented sample approach was not better than trimming the weights, this method is not discussed further.

## 4. BIAS IMPLICATIONS OF THE MODIFIED MITOFSKY-WAKSBERG METHOD

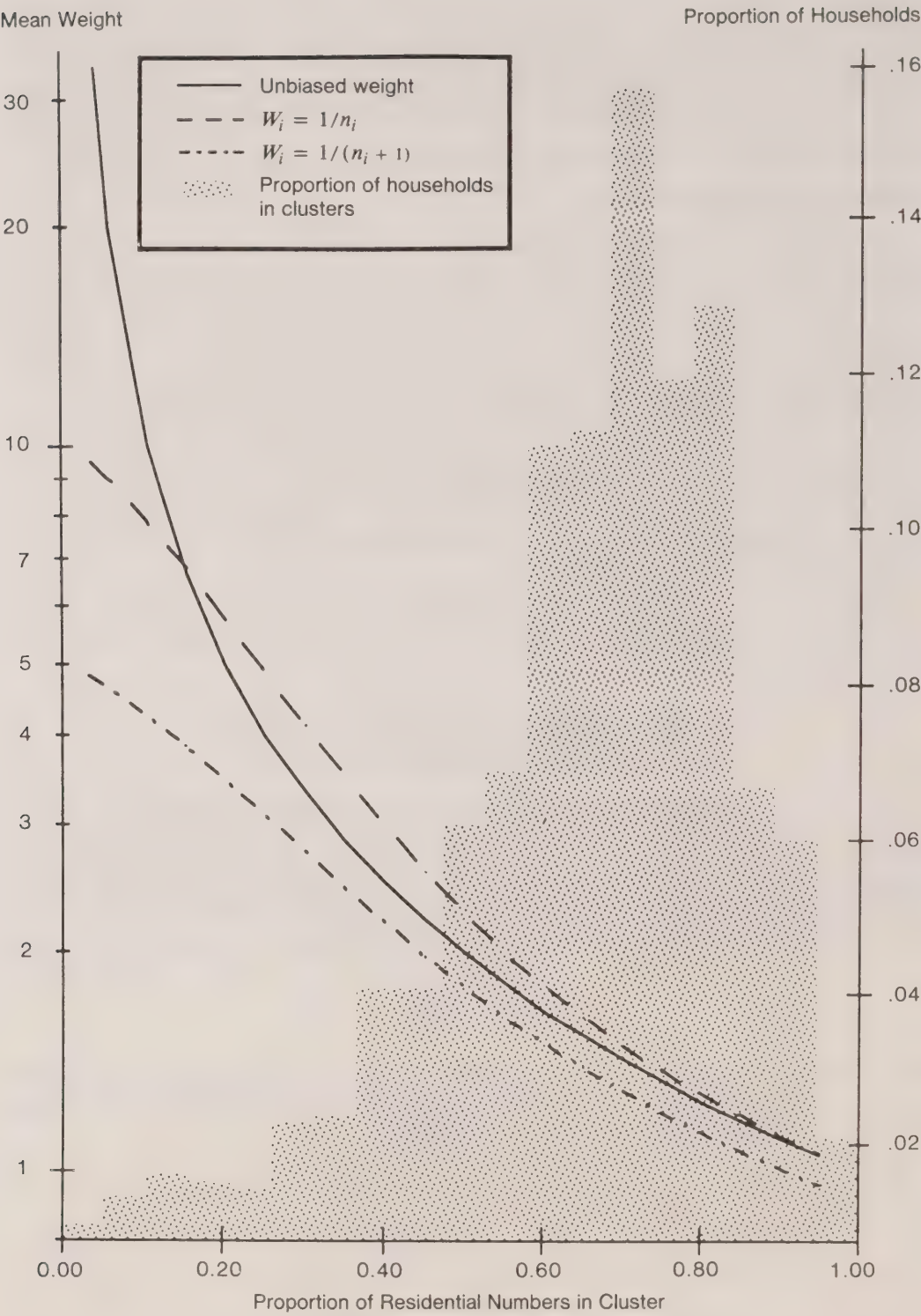
The increase in variance is just one of the consequences of using the modified Mitofsky-Waksberg method of sampling. Another important feature of the method is the bias in the resulting estimates. If a fixed sample size is selected in a cluster and no weight adjustment is made, the variance of the estimates are not increased but the bias has the potential of being very large.

The unbiased weight ( $W_u$ ) for the modified method is

$$W_u = \frac{100}{rN_i} \times \frac{100}{K},$$

where the terms are as defined above. The problem is that  $N_i$  is unknown and does not cancel with the second stage term, as it does in the standard Mitofsky-Waksberg method. Weights are therefore introduced in an effort to reduce the bias.

We refer to the estimator which uses a weight of  $n_i^{-1}$  as the natural estimator because  $n_i/K$  is an unbiased estimator of  $N_i/100$  in sampling from a binomial or hypergeometric distribution. (We use the weight of  $n_i^{-1}$  although the weight is actually  $K/100n_i$ . Since  $K/100$  is a constant, the relationship among the weights are not affected by using  $n_i^{-1}$ .) This weight appears to be the natural estimator despite the fact that  $n_i^{-1}$  is not unbiased for  $N_i^{-1}$  unless all 100 numbers are selected in a cluster. The bias of  $n_i^{-1}$  is discussed in literature; for example, see the discussion on stratification after sampling in Hansen, Hurwitz and Madow (1953). No simple unbiased estimator, certainly none of the form  $(n_i + t)^{-1}$ , is likely to exist for all possible cluster sample sizes.



**Figure 1.** Mean Weights of Estimators Conditional on the Proportion Residential with Shaded Histogram of Proportion of Households in Cluster



One of the ways to examine the potential bias is by comparing the expected value of the estimators (the mean weight using estimators of the form  $(n_i + t)^{-1}$ ) with the unbiased weight,  $W_u$ . Since both the unbiased weight and the expected value of the estimators are functions of  $N_i$ , we will begin by investigating these quantities conditioned on  $N_i$ .

Figure 1 shows the graph of the unbiased weight and the mean weights, using the estimators  $n_i^{-1}$  and  $(n_i + 1)^{-1}$ , when there are  $K = 10$  telephone numbers selected per cluster. The constant cluster sampling rate,  $r$ , has been omitted from all of the weights. A logarithmic scale has been used for the mean weights because of the range in  $W_u$ .

The graph clearly shows that the biggest differences between  $W_u$  and the mean weights for the two estimators are found when  $N_i/100$  is small. Once the residential density exceeds 20 percent when  $(n_i + 1)^{-1}$  is used, and 10 percent for  $n_i^{-1}$ , the differences are relatively minor. The graph shows that the weight  $(n_i + 1)^{-1}$  is always smaller than  $W_u$ , but this will not be true if poststratification is used. Poststratified weights are not used in the graph because poststratification really operates on the unconditional weights rather than the conditional weights shown here. The unconditional bias is addressed below.

The shaded histogram in the figure shows the distribution of households from Table 3. It has been overlaid to illustrate the fact that the large differences in weights occur in clusters which account for a very small fraction of the sampled households.

### Bias in Sample Size and Bias in Estimates

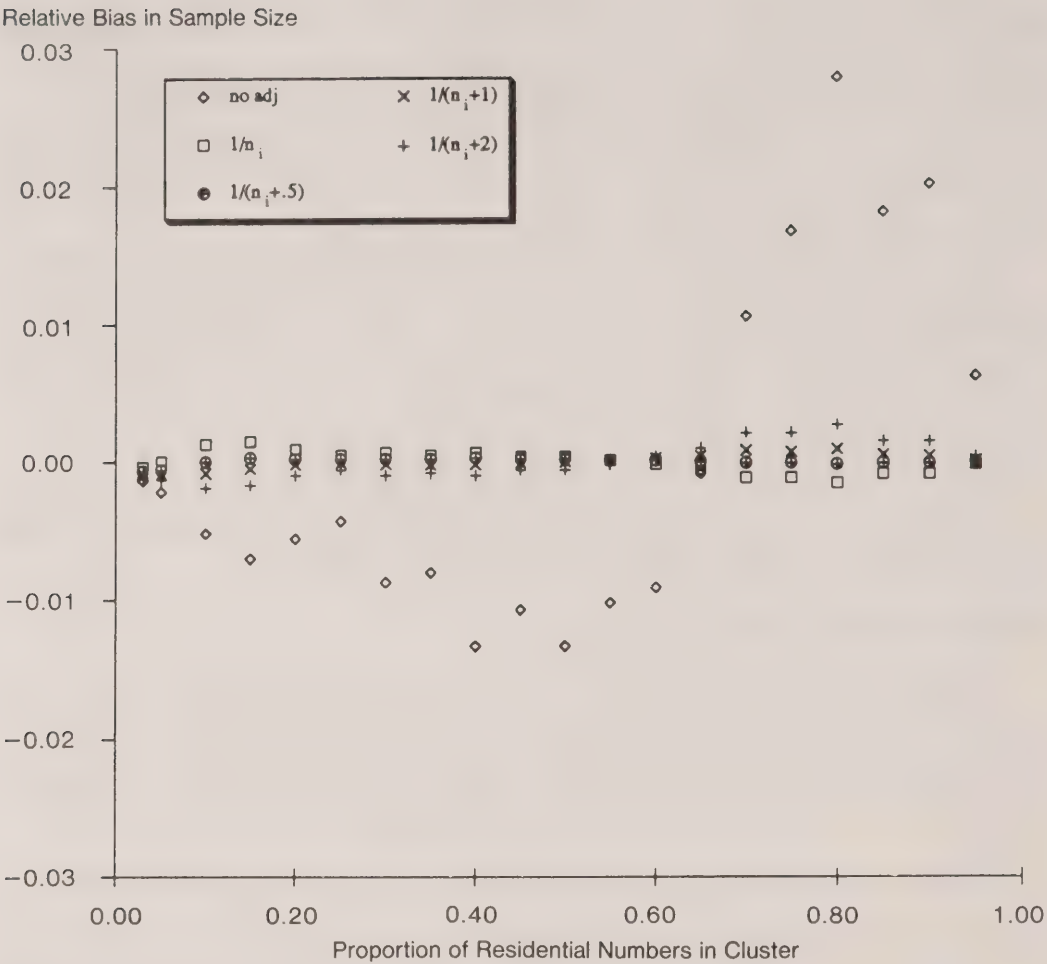
In nearly all RDD surveys, including those using the Mitofsky-Waksberg sample design, poststratification of the sample to known totals of persons or households is used. One of the prime reasons for using poststratification is to adjust the estimates to the levels associated with all persons, not just those in households with telephones. Massey and Botman discuss this and other benefits of poststratification in RDD surveys in Chapter 9 of Groves *et al.* (1988).

Regardless of the reasons for using it, poststratification results in estimates that are equal to known totals irrespective of the weights applied to the individual households. Since this bias, which can be considered as bias in sample size, is always zero, it is difficult to find a single statistic that measures unconditional bias directly. To attack this problem, we will examine the relative contribution to the bias in sample size over the range of household density values.

The following steps were taken to compute a measure of this contribution to bias in sample size. First, the different weighting functions or estimators were computed using the empirical household density shown in Table 3. Then, the estimates were poststratified to equal unity and the contribution to the total was computed for different values of  $N_i/100$ . Finally, the relative bias in sample size was defined as the difference between the contribution to the total from the particular estimator and the contribution from the total using  $W_u$  as the weight.

This measure thus takes into account both the difference in the weights for fixed values of  $N_i$  and the distribution of households across all the values of  $N_i$ . Thus, sampled households from clusters with values of  $N_i$  that are rare will not contribute heavily to the relative bias in sample size even if they are associated with large differences in weights.

To illustrate these computations, Figure 2 shows the relative bias in sample size for some estimators for samples of 30 numbers per cluster. One of the estimators uses the unadjusted weight, *i.e.*, the weight is a constant for all households regardless of the number of households identified in a cluster. The relative bias in sample size for the estimator with unadjusted weights is much larger than when other weights are used. The unadjusted weight has relative biases in sample size that range from about  $-2$  percent to  $+3$  percent.



**Figure 2.** Relative Bias from Sample Size for Samples of 30 Cluster

The size of the bias in the estimate of a characteristic is bounded by the size of the bias in sample size. In other words, the relative bias in the estimate can be no larger than the relative bias in the sample size. For almost all characteristics, this upper bound will not be attained. The upper bound is only attained when the characteristic and the residential density are perfectly correlated. Very high correlations are not likely in national samples, but might be more feasible in samples in restricted geographic areas.

It can be seen that there are patterns in the biases; for example, the unadjusted estimator is uniformly too low in low proportion residential clusters and too high in clusters with a high proportion of households. When there are differences in the characteristics between low and high density clusters, the biases can be quite serious. The bias in estimates resulting from using unadjusted weights can be seen for some characteristics in Table 1 in Cummings (1979). The biases are not very large, but appropriate weighting will effectively eliminate them.

In general, the relationship between the estimate and the number of households in a cluster will be unknown and inconsistent across all the characteristics to be estimated. Therefore, a reasonable practice is to choose an estimator that has a relative bias in sample size that is small across the range of values of  $N_i$ . If the relative bias in sample size for a set of the estimators is small, then the choice of estimators can be dictated by variance considerations.

### Biases Using Different Weights

The relative bias in sample size were computed using different estimators for samples of 5, 10, 30, and 60 telephone numbers per cluster. The relative bias in sample size is negligible for the cluster sample sizes of 30 and 60 numbers, except when the unadjusted weights are used. Any of the adjusted estimators could be used for cluster samples sizes of this size without incurring biases in the estimates. When 10 numbers are selected per cluster, all of the weights except the unadjusted one again perform reasonably well. The bias performance of  $(n_i + .5)^{-1}$  is especially encouraging.

For the smallest cluster size studied, 5 numbers per cluster, the potential for bias is somewhat greater. The natural weight,  $n_i^{-1}$ , has a somewhat lower bias in sample size than the weight  $(n_i + .5)^{-1}$ , but the difference is not very large. The relative bias in sample size for both of these weights is always less than one percent. For residential densities between about 45 percent and 80 percent the bias is positive and elsewhere it is negative. This pattern might be problematic only for the few characteristics that are very highly correlated with residential density.

### Biases with Trimmed Weights

The introduction of trimming can produce significant biases, depending on the relationship between the characteristics being estimated and the weights which are being trimmed. In some applications, the bias associated with trimming may limit the amount of trimming that can be applied and, hence, its usefulness for variance reduction.

The relative bias in sample size was also computed for cluster samples of 10, 30 and 60 numbers and the weights trimmed at about 3 times the mean weight. The trimming for samples of 10 numbers per cluster was done at a factor of 2 rather than 3 as described previously.

The difference between the relative bias in sample size for the trimmed and untrimmed weights is largely inconsequential for all cluster sample sizes and most values of  $N_i/100$ . The only noticeable difference occurred when the residential density is under about 10 to 15 percent. There is a slightly greater potential for bias in these regions. However, the relative bias in sample size for the trimmed weights is still much less than one percent at all residential density values.

## 5. CONCLUSION

The standard Mitofsky-Waksberg method is an effective method of producing a self-weighting, RDD sample of fixed size. However, the sequential monitoring of the number of cases per cluster is an awkward operational feature of this method. One of the consequences of the sequential monitoring of caseloads by cluster is that it is difficult to complete data collection in a tight time frame. The data collection period has to be flexible enough to allow for obtaining the appropriate number of cases in each cluster. The more extensive data collection period and the monitoring of caseloads also result in increasing costs. Another problem with the sequential operations is that the requirement for frequent monitoring of the caseloads can lead to frustration arising from complications of combining sample selection and data collection operations.



The modified Mitofsky-Waksberg approach eliminates the sequential nature of the design and, with it, the need to monitor the work by cluster. A fixed number of telephone numbers are assigned to each sampled cluster in the modified method. Therefore, the costs associated with monitoring caseloads and a longer data collection period are not incurred. However, the modified Mitofsky-Waksberg method does introduce new components of bias and variance into the estimates. These statistical concerns should be addressed before the modified approach is used.

Specific recommendations on when the standard or modified Mitofsky-Waksberg method should be used are difficult to formulate since they depend upon circumstances which vary from survey to survey. Guidelines for choosing between the methods are suggested below.

A simple rule is that for surveys which require either very tight controls on sample size or a nearly self-weighting sample, then the standard Mitofsky-Waksberg approach is advisable. Even though the sample size in the modified method can be estimated relatively precisely, some variation, especially because of uncertainty of the nonresponse rates, can be expected. A self-weighting sample, which is not achieved when the modified Mitofsky-Waksberg method is used, also has some advantages in simplifying standard statistical analysis.

Since the costs for standard and modified methods are different, it would be very useful to have cost-variance models to help evaluate the two methods. Unfortunately, the differences in costs of the standard and modified methods are not easy to quantify. In fact, the lack of reasonable cost models is a major and pervasive problem that limits the ability to establish optimal survey design.

Because of lack of reasonable cost-variance models, we suggest some conditions in which one approach might be favored over another. One of the conditions that favors the modified approach is a relatively brief interview length. As the interview becomes longer, the cost savings associated with the modified method is likely to become smaller relative to the increases in variances of the estimates.

The length of the interview is particularly important for surveys which screen households to find those with particular characteristics. For example, some RDD surveys screen households and only interview if a member is in a particular target group. In these situations, the screening interview is often very brief. The modified Mitofsky-Waksberg approach may be very beneficial. Surveys in which households are screened also tend to have large cluster sample sizes, and this improves the performance of the modified procedure. When 10 or more numbers are selected per cluster (equivalent to about 6 households per cluster), the biases in the estimates under the modified Mitofsky-Waksberg approach are virtually inconsequential and the increases in variance with trimming are only about 10 percent. Samples of 10 or more numbers per cluster are frequently acceptable for screening purpose although such large cluster sizes are typically inefficient for the interview sample, even when the intraclass correlation is small.

Based on these factors, a general guideline is that the modified Mitofsky-Waksberg method can be recommended when households within the clusters must be screened. More specifically, the modified method with trimmed weights should be considered if the following conditions exist: (1) Ten or more numbers are sampled per cluster; and (2) the total cost for the modified method is at least 10 percent less than the standard method, or the data collection period is relatively short. If both of these conditions are not met, then the choice between methods must be made on evaluations of other survey requirements.

When the cluster sample size is less than 10, the bias and variance arising from the use of the modified Mitofsky-Waksberg method are more serious concerns. Any characteristics correlated with the proportion of residential numbers in a cluster could be affected with a cluster sample size this small. Also, the variance of the estimates with the modified method will be

20 to 30 percent larger than with the standard method since trimming is not very effective with small sample size. Therefore, in most surveys with sample sizes of less than 10 numbers per cluster, the problems of implementing the standard method should be quite serious before a decision is made to abandon it and use the modified method.

### ACKNOWLEDGEMENT

The author would like to thank the referees. Their comments were very helpful in improving the presentation of the paper.

### REFERENCES

- CUMMINGS, K.M. (1979). Random digit dialing: A sampling technique for telephone surveys. *Public Opinion Quarterly*, 233-244.
- DREW, J.D., DICK, P., and SWITZER, K. (1989). Development and testing of telephone survey methods for household surveys at Statistics Canada. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 120-127.
- GROVES, R.M., BIEMER, P.P., LYBERG, L.E., MASSEY, J.T., NICHOLLS, W.L., and WAKSBERG, J. (editors) (1988). *Telephone Survey Methodology*. New York: John Wiley and Sons.
- HANSEN, M.H., HURWITZ, W.N., and MADOW, W.G. (1953). *Sample Survey Methods and Theory*, 2. New York: John Wiley and Sons, 138-139.
- KISH, L.(1965). *Survey Sampling*. New York: John Wiley and Sons, 429-430.
- LEPKOWSKI, J.M., and GROVES, R.M. (1986). A two phase probability proportional to size design for telephone sampling. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 73-98.
- PIEKARSKI, L. (1990). Working block density declines. *The Frame*, a publication of Survey Sampling Inc.
- POTTHOFF, R.F. (1987). Generalizations of the Mitofsky-Waksberg technique for random digit dialing. *Journal of the American Statistical Association*, 82, 409-418.
- SUDMAN, S. (1973). The uses of telephone directories for survey sampling. *Journal of Marketing Research*, 10, 204-207.
- TUCKER, C. (1989). Characteristics of commercial residential telephone lists and dual frame designs. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 128-137.
- WAKSBERG, J. (1984). Efficiency of alternative methods of establishing cluster sizes in RDD sampling. Unpublished Westat Inc. memorandum.
- WAKSBERG, J. (1978). Sampling methods for random digit dialing. *Journal of the American Statistical Association*, 73, 40-46.
- WAKSBERG, J. (1973). The effect of stratification with differential sampling rates on attributes of subjects of the population. *Proceedings of the Social Statistics Section, American Statistical Association*.





# The Blaise System for Integrated Survey Processing

JELKE G. BETHLEHEM and WOUTER J. KELLER<sup>1</sup>

## ABSTRACT

Application of recent developments in computer technology allow national statistical offices to produce high quality statistics in an efficient way. At the Netherlands Central Bureau of Statistics (CBS) an increasing use is made of microcomputers in all steps of the statistical production process. This paper discusses the role of software and hardware in data collection, data editing, tabulation, and analysis. To avoid the negative effects of uncontrolled de-centralized data processing, the importance of integration is stressed. This makes the statistical production process easier to manage, and moreover it increases its efficiency. The Blaise System, developed by the CBS, is discussed as a data processing tool that encourages integration. Using a description of the survey questionnaire, this system is able to automatically generate various computer programs for data collection (CAPI or CATI), or data entry and data editing (CADI). The system can also create interfaces to other packages. Particularly, the link between Blaise and the internally developed packages Bascula (for weighting) and Abacus (for tabulation) is described. In this way the Blaise System controls and co-ordinates, and therefore integrates, a large part of the survey process.

**KEY WORDS:** Integration; Survey processing; CAPI; CATI; Microcomputers; Decentralization; Standardization.

## 1. INTRODUCTION

The Netherlands Central Bureau of Statistics (CBS) makes an increasing use of microcomputers in survey data processing. The introduction of microcomputers has a considerable impact on the way the work of the statistical office is carried out. Subject matter statisticians become increasingly aware of the potential of the new technology, and consequently use it more and more in their daily work.

This paper discusses the role of the new automation technology in data collection, data editing, tabulation, and analysis. We will stress the importance of standardization and integration. These working policies have three advantages: they enable us to avoid the negative effects of uncontrolled de-centralized data processing, they make the statistical production process easier to manage, and they increase efficiency.

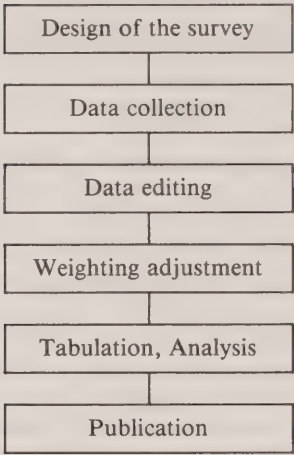
The Blaise System, developed by the CBS, is discussed as the backbone of an integrated survey processing system. On the one hand, the power of this system lies in the consistency it enforces in the various steps of data collection and data processing. On the other hand, it also promotes standardization between different departments. Since all departments use the same software for processing their surveys, everybody speaks the same "language", and so exchange of information between departments is easier and less error prone.

## 2. THE STATISTICAL PRODUCTION PROCESS

National statistical offices collect data on persons, households and establishments and transform this information into useful statistics. Production of statistical information is often a complex, costly and time-consuming process. This section describes the various steps the

<sup>1</sup> Jelke G. Bethlehem and Wouter J. Keller, Netherlands Central Bureau of Statistics, Automation Department, P.O. Box 959, 2270 AZ Voorburg, The Netherlands.

**Table 1**  
The statistical production process



statistical office has to go through, the problems that it may encounter, and the decisions it has to make. An overview of the process is given in table 1.

The first step is, of course, the design of the survey, in which the statistician specifies the population to be investigated, the data to be collected, and the characteristics to be estimated. Since statistical offices collect most data by means of (sample) surveys, a questionnaire has to be defined, containing the questions to be asked of the respondents. This questionnaire is the first practical description of the data to be collected. Furthermore, in the case of a sample survey, the statistician also has to specify a sampling design, and he must see to it that the sample is selected properly.

The second step in the process is *data collection*. Traditionally, in many surveys the questionnaires are completed in face-to-face interviews: interviewers visit respondents, ask questions, and record the answers on (paper) forms. The quality of the collected data tends to be good. However, since it typically requires a large number of interviewers, who may all have to do much travelling, it can be expensive and time-consuming. Therefore telephone interviewing is sometimes used as an alternative. The interviewers call the respondents from the statistical office, and thus no more travelling is necessary. However, telephone interviewing is not always feasible: only connected people can be contacted, and the questionnaire should not be too long nor too complicated. A mail survey is cheaper still: no interviewers at all are needed. Questionnaires are mailed to potential respondents with the request to return the completed forms. Although reminders can be sent, the persuasive power of the interviewer is lacking, and therefore response tends to be lower in this type of survey, and so does the quality of collected data.

If the data are collected by means of paper forms, completed questionnaires have to undergo extensive treatment. In order to produce high quality statistics, it is vital to remove any errors. This step is called *data editing*. Three types of errors can be distinguished: A *range error* occurs if a given answer is outside the valid set of answers, e.g. an age of 348 years. A *consistency error* indicates an inconsistency in the answers to a set of questions. An age of 8 years may be valid, and a marital status “married” is not uncommon, but if both answers are given by the same person, at least in the Netherlands, there is something definitely wrong. The third

type of error is the *routing error*. This type of error occurs if the interviewer or the respondent fails to follow the specified branch or skip instructions, *i.e.* the route through the questionnaire is incorrect: irrelevant questions are answered, or relevant questions are left unanswered.

Detected errors have to be corrected, but this can be very difficult if it has to be done afterwards, at the office. In many cases, particularly for household surveys, respondents cannot be contacted again, so other ways have to be found to do something about the problem. Sometimes it is possible to determine a reasonable approximation of a correct value by means of an imputation technique, but in other cases an incorrect value is replaced by the special code indicating the value is "unknown".

In addition to data editing, another activity is sometimes carried out during this stage of the production process: *coding of open answers*. A typical example is the question about the occupation of the respondent. Questions are easiest to process if a respondent selects one possibility from a list of pre-coded answers. However, for a question like occupation this set of pre-coded answers would be very long, and thus it would be very hard for the respondent to select the proper answer. This problem is avoided by letting the respondent formulate his own answer, and then literally copying the answer on the form. To enable analysis of this type of information, answers must be classified afterwards. This is a time-consuming and costly job, which must be carried out by experienced subject-matter specialists.

After data editing, the result is a "clean" file, *i.e.* a file without errors. However, this file is not yet ready for tabulation and analysis. In the first place, the sample is sometimes selected with unequal probabilities, *e.g.* establishments are selected with probabilities proportional to their size. The reason is that a clever choice of selection probabilities makes it possible to produce more accurate estimates of population parameters, but only in combination with an estimation procedure which corrects for this inequality. In the second place, representativity may be affected by nonresponse, *i.e.* for some elements in the sample the required information is not obtained. If nonrespondents behave differently with respect to the population characteristics to be investigated, the results will be biased.

In order to correct for unequal selection probabilities and nonresponse, a *weighting adjustment* procedure is often carried out. Every record is assigned some weight. These weights are computed in such a way that the weighted sample distribution of characteristics like sex, age, marital status and area reflects the known distribution of these characteristics in the population.

In the case of item non-response, *i.e.* answers are missing on some questions, and not all questions, an *imputation procedure* can also be carried out. Using some kind of model, an estimate for a missing value is computed and substituted in the record.

Finally, we have a clean file which is ready for *analysis*. The first step in the analysis phase will nearly always be tabulation of the basic characteristics. Constructing a table is not as simple as it may look at first sight. The composition of rows and columns (often built from a number of variables), the quantities displayed in cells (counts, means, percentages), the way in which percentages are computed, treatment of multiple-response variables, the position of totals and subtotals, and many other things, can make life very difficult.

Many statistical offices also carry out analysis on their data in order to reveal the underlying structures, and thus to gain insight in the data. Information obtained in this way may improve a later survey, and thus improve quality or reduce costs.

The results of the analysis will be *published* in some kind of report. Usually it will contain tables and graphs. It is important to present the statistical information in such a way that the proper "message" is conveyed. Graphs, tables and text should be simple and clear. Particular attention should be paid to graphs, because a visually ambiguous or confusing graph will quite easily lead to wrong interpretation.



### 3. THE NEED FOR INTEGRATION

The computer has always been important in statistical information processing. In the beginning to computer was only used for activities like sorting, counting and tabulation. In the sixties and seventies, with the emergence of mainframes and statistical packages, it became possible to carry out extensive analysis. The computer was also increasingly used for data editing, weighting adjustment and imputation. The use of computers for data collection is more recent. This first occurred for telephone interviewing (CATI). In the last decade, the advent of the small laptop computers has made it possible for interviewers to take the computer with them to the homes of the respondents. This way of computer assisted face-to-face interviewing is denoted by CAPI.

It will be clear that the computer is used for more and more activities. Hardware and software are available for nearly every step in the production process. Also an increasing number of people are making use of the automation tools. At first, only the computer specialists had access to “their” machines, but now statisticians and subject matter experts have become computer-oriented, and therefore make increasing demands for suitable software and hardware to do their jobs. Simple and straightforward electronic data processing can, and is, now carried out by the subject matter departments themselves, leaving design and maintenance of complex information systems to be carried out by the computer specialists of the automation department. As a consequence of these developments the work of the statisticians and subject-matter experts have changed. They used to be specialists in their own (narrow) field, but now they have acquired more general knowledge and experience in a much broader field containing subject matter aspects, statistical methodology and computer processing. So the specialists have vanished, and a new group with general knowledge of all aspects of survey processing has emerged.

Automation of the statistical production process is nice, but one should be aware of the dangers. Although application of computers promises increased efficiency and quality, an uncontrolled and unco-ordinated use of the new technology may easily lead to chaos, and hence to less productivity. Factors affecting the efficiency of the statistical production process are:

– **Different departments are involved.**

Many people deal with the information: respondents fill in forms, subject-matter specialists check forms and correct errors, data typists enter the data in the computer, and programmers construct editing programs. Transfer of material from one person/department to another can be a source of error, misunderstanding and delay.

– **Different computer systems are involved.**

Various data processing activities may be carried out on different computer systems. Transfer of files causes delay, and incorrect specification and documentation may produce errors.

– **Repeated specification of the data.**

In almost every step of the process, the structure of the data must be specified. The particular system or department has to know about the data: What is the meaning of the variables? Which values are permitted? Are there any constraints on the routing? Which relationships between variables have to be checked? Although essentially the same, the form of specification may be completely different for every step. Every system uses its own “language”. The first specification is the questionnaire itself. Another specification may be needed for data-entry, and yet another for the checking program, for tabulation and analysis, *etc.* It is clear that this is not the most efficient way to deal with the information.

The CBS solution to these problems is *integration*. In this context, integration has three different aspects: integration of work, hardware, and software. Let us first have a look at integration of the work.

Traditional data processing consist of what we call *macro cycles*. All survey data as a whole goes through cycles: from one department to another, and from one computer system to another. First the paper forms are cleaned manually by the subject matter department, then data on the forms are entered by the data entry department, next the files are transferred to a mainframe computer system. A program checks the data for consistency, detected errors are printed on lists that are send back to the subject matter department for corrections. This process of data entering and data editing has to be repeated a number of times before the data can considered to be "clean".

The idea behind integration of work is that the macro cycles should be replaced by *micro cycles*. Not the whole data file, but instead only one record at a time should cycle around. Micro cycles means that cycling should take place within one computer system, and that this should be controlled by one department. Going from macro cycles to micro cycles comes down to concentrating all data processing activities in one department, and that is the subject-matter department. Since the subject-matter statisticians are the ones with most knowledge about the area covered by a survey, they are best equipt to deal with the data, to solve problems, and to produce high quality statistics. Of course, they need proper instruments to do their job, *i.e.* powerful and user-friendly software and hardware.

The idea that automation of data processing activities should be carried out exclusively by computer specialists is out of date. More and more the subject-matter statisticians become aware of the possibilities and usefulness of the computer for their own work. So the time has come for subject-matter departments to take simple and straightforward survey data processing into their own hands. Of course, the automation department is responsible for providing the proper automation infrastructure. And this department stays in charge of design and maintenance of complex information systems.

The second aspect of integration is integration of hardware. The idea is to concentrate work on one type of computer as much as possible. Taking into account that a large number of inexperienced statisticians will have to use the computer, the obvious choice is the microcomputer. Microcomputers offer user-friendliness at a relative low price, and moreover, there is an abundance of useful software.

Being aware of the fact that statistical offices process huge quantities of data, one may wonder whether microcomputers have the capacity to carry out all work, and indeed can take over from the large workhorses, the mainframes. To be able to answer this question, it is useful to distinguish between two kind of data processing activities. In the first place there are record oriented activities. These are activities for which only one record at a time is needed. Examples of record oriented activities are data entry and data editing. Record oriented activities are generally very well suited for interactive processing. In the second place, there are file oriented activities. These activities can only be carried out properly if the whole file is available. Examples are the computation of weights and tabulation. Because of their size, file oriented activities are often processed in a batch-wise fashion.

The viewpoint of a few years ago was that record oriented activies could be carried out on microcomputers but file oriented activites had to take place on mainframes. With the increasing power of microcomputers, attention is shifting in the direction of the microcomputer. At this moment, the policy of the CBS is that all record oriented activities have to be carried out on microcomputers and file oriented activities can in many cases (say, with data files of less than 50 megabytes) also be carried on microcomputers. However, for data storage and large batch jobs we still need mainframes.



The users of the computer environment should be confronted as little as possible with the mainframe. Therefore, the CBS is moving in the direction of front end/back end systems. The front end consists of microcomputers, and that is what the statisticians use to specify their problems. The back end is a mainframe or mini-computer, and is used bulk work, maybe even without the user knowing it. Particularly for database applications the client/server approach looks very promising. In this approach, the real database activities take place on a dedicated minicomputer, whereas the activities are specified, initiated and controlled by the microcomputers at the desks of the users.

#### 4. STANDARDIZATION

The CBS makes an increasing use of microcomputers (running under MS-DOS) in many steps of the statistical production process. On the one hand, this opens new ways towards efficient information processing, but on the other hand, it creates new problems that have to be dealt with. If every department is free to select and purchase its own type of computer and software, the automation infrastructure may easily get out of control, and turn into chaos. Departments will not talk the same "language" anymore, because they use different data formats and different software. It is clear that this calls for a strong policy on standardization. The CBS has adopted such a policy, and in practice it means that there are only one or two software packages available for a particular task.

Another advantage of standardization is that it limits the amount of training that has to be provided for the users. In order to cope with the problem of training a large number of new microcomputer users, the CBS runs an average of 50 one-day courses per month (occupying three fully equipped lecture rooms every working day).

Attention should also be paid to the way in which the microcomputers are used in the organization. Distribution of a lot of stand-alone microcomputers may seem a simple solution, but there are also problems that have to be solved. In the first place, it is very easy to copy (confidential) data files on local hard disks, so we have a data security problem. Furthermore, activities like making back-ups and archiving are often neglected by the users in the subject-matter departments. Also communication between departments (*e.g.* sharing data files) is only possible by exchanging floppy disks. Finally, distribution of new releases of software packages, including their documentation, is often cumbersome in large organizations with a lot of stand-alone microcomputers.

To avoid the above mentioned problems, the CBS has installed approximately 60 local area networks (LANs). Every department has its own LAN. Ten to sixty microcomputers are connected to a high-end 386-based fileserver with a storage capacity of up to 600 Megabytes. In this environment there are in total nearly 2,300 microcomputers, half of them based on the Intel 386SX micro-processor. Security is guaranteed by means of password protection in a login-procedure, by encryption, and by using floppy-less workstations (of the 2,300 microcomputers only 60 have a floppy or hard disk drive). Archiving and backing-up the LANs is carried out in a centralized way by the automation department. A full backup of more than 15 Gigabytes is carried out every night. It is clear that version control and updating software can more easily be realized in such an environment. Distribution and installation of new software releases on a LAN is easy, since, with one command one can upload the new version to all file servers. All software licenses are based on concurrent usage, which is checked by home-made software.

The role of microcomputers in the statistical production process is growing, but for the time being, there are still applications (like the use of large databases) that need mainframe or minicomputer systems. In this environment, the CBS has adopted Oracle as the standard



database system. Development of a database application is preferably carried out on a microcomputer, whereas actually running it takes place on a mini computer. Recently, the CBS realized a client/server architecture based on a distributed database system. Microcomputers in the network serve as front ends and the minicomputers as back ends.

So, as the use of the data processing instruments is brought closer to the subject-matter specialists at the departments (de-centralization), standardization and coordination of the work environment of the subject-matter users demands strong centralization. More details about the automation infrastructure can be found in Keller, Metz and Bethlehem (1990).

## 5. INTEGRATION OF THE SURVEY PROCESS

The previous section discussed the need for integration in the survey process. Particular attention was paid to concentrating the work in subject-matter departments, and standardization of the hardware and software instruments. But standardization of software is not enough. The efficiency of the production process can be increased even more by integrating the required standard software into one system. This section describes how such an integrated system for survey processing is implemented at the CBS.

An integrated system for survey processing should be based on a powerful language for the specification of questionnaires. This specification is the "knowledge base", containing all knowledge about the questionnaire and the data. The system should be able to exploit this knowledge, *i.e.* it must be able to automatically generate all required data processing applications. On the one hand it means the automatic generation of software for data collection, data entry and data editing, and on the other hand the automatic generation interfaces for other data processing software, *e.g.* for tabulation and analysis. In this way repeated data specification is no longer necessary, and consistency is enforced in all data processing steps.

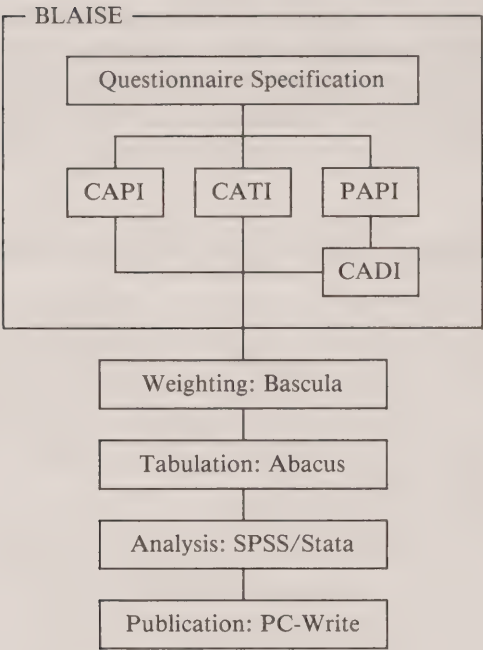
The backbone of the integrated survey processing system developed by the CBS is the Blaise System. In the design phase of the survey, the questionnaire is specified in the Blaise language. And it is this specification that is used throughout the whole survey process to extract the information necessary to carry out the various data processing steps. Table 2 summarizes the integrated system for survey processing.

The Blaise System can produce three kinds of programs: CADI, CAPI and CATI programs. CADI stands for Computer Assisted Data Input. It integrates data entry and data editing by offering an interactive environment for processing paper questionnaire forms. The Blaise System can also produce the software required to carry out CAPI or CATI interviewing. The Blaise System is discussed in more detail in section 6.

Whatever form of data collection is used, the result will be a "clean" data file, *i.e.* a file in which no more errors can be detected. The next step in the process will often be the computation of adjustment weights. The program Bascula will take care of this. It is able to read the Blaise data files directly, and extract the information about the variables, *i.e.* the meta-information, from the Blaise specification. Running Bascula will cause an extra variable to be added to the data file containing the adjustment weight for each case. More about Bascula can be found in section 7.

Now the file is ready for tabulation, and for that, the integrated system offers the program Abacus. This program is also able to read and understand the data files created in the previous step of the process. See section 8 for details. Tabulation may be followed by a more extensive analysis of the data. For that purpose the Blaise System can generate interfaces for the statistical packages SPSS and Stata. More about this in section 9.

**Table 2**  
Integrated survey processing



Finally, a publication will be prepared using the standard wordprocessor PC-Write. Since this wordprocessor runs on the same computer system as the other software, it is easy to import generated tables and results of statistical analysis into the text.

**6. THE BLAISE SYSTEM**

The Blaise System was developed by the CBS, and it derives its name from the famous French theologian and mathematician Blaise Pascal (1623-1662). The basis of the Blaise System is the Blaise language, which is used to create a formal specification of the structure and contents of the questionnaire. The Blaise language has its roots, in large part, in the programming language Pascal.

The Blaise System runs on microcomputers (or networks of microcomputers) under MS-DOS. It is the backbone of the integrated survey processing system, and as it is intended to be used by the people of the subject-matter departments, one need not be a computer expert to use the Blaise System. The design goal of the system was to provide subject-matter experts with a powerful but user-friendly tool that enables them to input their knowledge about a survey into the system, and to take care of all subsequent data processing steps.

In the Blaise philosophy, the first step in carrying out a survey is to design a questionnaire in the Blaise language. Such a specification of the questionnaire contains more information than a traditional paper questionnaire. It not only describes questions, possible answers, and conditions on the route through the questionnaire, but also relationships between answers that have to be checked.

The Blaise System can produce programs for CADI, CAPI or CATI. A CADI program is an intelligent and interactive system for data entry and data editing of data collected by means of paper forms. The subject-matter specialist works through a number of forms with a microcomputer, processing them one-by-one. He enters answers to questions at the proper places and, after completion of the form, he activates the check option to test routing and consistency. Detected errors are reported and explained on the screen. Errors can be corrected by consulting the form or calling the supplier of the information. After elimination of all errors, a clean record is written to a file.

The CADI program can also be used for a different way of data processing not mentioned thus far. Sometimes statistical offices do not carry out their own data collection, but they have to create statistics using data files that were generated elsewhere, outside the statistical office. In these cases the data still has to be checked. The Blaise System has a facility to import this kind of data files. With a CADI program, an integral check can be carried out on all records in a batch-wise version. Thus the records are assigned either the status "clean" or "dirty". And the dirty records can be corrected interactively, again with the CADI program.

A CAPI/CATI program can be used for computer assisted interviewing. The paper questionnaire form is replaced by a computer program containing the questions to be asked. This computer program is in control of the interview. It determines the proper next question to be asked, and checks the answers as soon as they have been entered. In the case of CAPI, the interviewing program is loaded into a laptop computer, and the interviewer takes this computer to the homes of the respondents. In the case of CATI, the program is in a desktop computer. The interviewer calls the respondents from a central unit, and carries out the interview by telephone.

The generation of a Blaise CADI/CAPI/CATI proceeds in a number of steps. First, a text editor is used to enter the Blaise specification of the questionnaire, after which it is checked for syntax errors. Detected errors must be corrected, and to do that the system returns to the text editor and places the cursor on the approximate location of the error. After correction, the specification is checked again. If no errors are detected, the specification is transformed into Pascal source code, which in turn is compiled into an executable program.

The Blaise language must serve two somewhat conflicting purposes. On the one hand it must be powerful enough to be able to deal with all kinds of large and complex surveys, and on the other, Blaise questionnaire specifications must be readable enough, for use by subject matter specialists. In fact, a Blaise questionnaire must be self-documenting, i.e. it is the basic description of the survey which can be used by all people involved. Table 3 gives an example of a simple questionnaire in Blaise.

The first part of the questionnaire specification is the QUEST section, containing the definition of all questions that can be asked. A question consists of an identifying name (for internal use in the questionnaire), the text of the question as presented to the respondents, and a specification of valid answers. The next part of this sample Blaise questionnaire is the ROUTE section. It describes under which conditions, and in which order the questions have to be asked. Consistency checks are specified in the CHECK section.

The description above does not exhaust the power of the Blaise language. An overview of the Blaise language can be found in Bethlehem *et al.* (1989b), and more details in Bethlehem *et al.* (1989c).

The Blaise System contains a module for *interactive coding*, thereby providing the possibility of integrating coding either in the data collection phase or in the data entry and data editing phase. The module contains two different tools. The first tool implements a hierarchical approach to coding. Coding of an answer starts by entering the first digit of the code by selecting



**Table 3**  
A simple Blaise questionnaire

---

QUESTIONNAIRE Work “The Work Survey”;	
QUEST	
SeqNum	“Sequence number of the interview?”: 1..1000 (KEY);
Age	“What is your age?”: 0..99;
Sex	“Are you male or female?”: (Male, Female);
MarStat	“What is your marital status?”: (Married “Married”, NotMar “Not married”)
Job	“Do you have a job?”: (Yes, No);
JobDes	“What kind of job do you have?”: STRING[20];
Income	“What is your yearly income?”: (Less20 “Less than 20,000”, Upto40 “Between 20,000 and 40,000”, More40 “More than 40,000”);
Travel	“How do you usually travel to your work?”: SET [3] OF (Walking “Walking”, Bicycle “By bicycle”, Car “By car or motorcycle”, PubTrans “By bus, tram, train or metro”, Other “Other means of transport”);
OthTrans	“What other means of transport?”: STRING[20];
ROUTE	
SeqNum; Age; Sex; MarStat; Job;	
IF Job = Yes THEN	
JobDes; Income; Travel;	
IF Other in Travel THEN OthTrans ENDIF	
ENDIF	
CHECK	
IF Age < 15 “Respondent is younger than 15” THEN	
MarStat = NotMar “he/she is too young to be married!”	
ENDIF	
ENDQUESTIONNAIRE.	

---

the proper category from a menu. After the user enters a digit, the program presents a subsequent menu containing a refinement of the previously selected category. So the description becomes more and more detailed until the final digit is reached. The second tool consists of a dictionary approach to coding. It tries to locate an entered description in an alphabetically ordered list. If the description is not found, the list is displayed, starting at a point as close as possible to the entered description. The list can be made so that almost any description, including permutations, is present. The advantage of this method is that it is simple, fast and controllable. Both coding tools can be used simultaneously.

## 7. BASCULA

The clean file with sample survey data produced by the Blaise System is usually not ready yet for making inference about the population from which the sample has been drawn. The problem is that the data do not constitute a representative sample, and so some adjustment procedure has to be carried out.

In order to account for unequal selection probabilities and nonresponse, one often has to compute adjustment weights. Post-stratification is a well-known technique. Every record is assigned some weight, and these weights are computed in such a way that the weighted sample distribution of characteristics like sex, age, marital status, and area reflects the known distribution of these characteristics in the population. Two major problems can make application of post-stratification difficult: empty strata and lack of adequate population information. Research has been carried out at the CBS in order to improve weighting techniques. The result was a new general method for weighting, in which weights are obtained from a linear model which relates the target variables of a survey to auxiliary variables. Post-stratification is a special case of this method. Because of the generality of the method, different weighting schemes can be applied that take advantage of the available population information as much as possible, and at the same time avoid the above mentioned problems. See Bethlehem and Keller (1987) for more details.

Bascula is a general weighting program, running on microcomputers under MS-DOS. It combines several weighting techniques. In the first place, traditional post-stratification can be carried out. And if the number of empty strata is small, one can instruct the program to collapse (*i.e.* combine) these strata with neighbouring strata. In the case of many empty strata, or lack of sufficient population information, Bascula can carry out the linear weighting technique described above or apply iterative proportional fitting (also called multiplicative weighting, or raking ratio estimation). The resulting weights can either be added to the data file, or be stored in a separate file.

Bascula is able to read the Blaise data file directly, and also extracts the required information about the variables from the Blaise specification. The information about the population has to be provided by the user. The program is menu-driven, making it user-friendly. It will carry out a complete post-stratification if possible. If not, the user has to decide either to carry out linear or multiplicative weighting.

Presently, Bascula can only be used on a microcomputer. In the future, a back end will be developed that will run on our mainframe environment. Bascula was particularly developed for use in social and demographic surveys, where post-stratification is combined with relatively simple estimation procedures. For use in economic surveys, different software will be developed. This software will concentrate on stratified sampling designs in combination with more complex estimators (ratio and regression estimation).

## 8. ABACUS

Tabulation is one of the basic activities in the statistical production process, and it was one of the first to be automated. Many tabulation packages have already been developed in the world, but many are not very user-friendly. This is partly caused by the fact that proper generation of a complex table needs a lot of parameters to be specified: the variables to be used in the various dimensions (rows, columns, layers), whether to concatenate variables (display all values of a variable, followed by all values of another variable) or to nest variables (display for every value of one variable all possible values of another variable) within a dimension, the

displayed cell quantity (counts, percentages, totals, averages), whether to display totals and subtotals, and many layout features. To be able to cope with all these parameters, traditional packages have control languages to specify tables, and these languages are often not very easy to learn and to use.

Abacus is a tabulation package, running on microcomputers under MS-DOS. While Abacus may be seen as yet another tabulation package, it was developed with very specific design goals. In the first place, no control language is used to specify a table. The program is menu-driven instead. The user designs his table in an interactive, simple, and intuitive way, without having to know about any control language. In the second place, Abacus can directly read the data file created by the Blaise System, as well as Ascii files. The meta-information, *i.e.* the information about the variables in the file, can be generated by the Blaise System, or can (in the case of separate Ascii files) be entered interactively by the user. Thirdly, the program can produce camera-ready tables.

Another striking property of Abacus is its speed. A table produced by SPSS-Tables in 3 minutes was generated by Abacus in about 6 seconds (all timings based on the same 386SX based microcomputer). The reason for this is that the Abacus program is rather small, so it can use a large part of the memory as a working area which allows for a table of up to 90,000 cells.

Tables produced by Abacus can have up to three dimensions (layers, rows and columns). Every dimension can hold up to 10 variables, which may be nested or concatenated. In the example in table 4, the column variables "Employment" and "Sex" are concatenated while in the row dimension the variables "Region" and "Town" are nested. In this example no variable has been placed in the layer dimension.

This table contains simple counts, but Abacus can also calculate totals of quantitative variables, make percentages tables, and averages tables. It is also possible to have more than one (up to 10) items in the cells of the table. In that case, the user has to decide to put each item in a separate row, column or layer. If the data has been collected by means of a sample survey, Abacus can accommodate weighted data, using the weights that are, for example, computed by Bascula. The only thing the user has to do is to specify the variable containing the weights.

Table 4  
An example of a two-dimensional table

Number of Records	The population of Samplonia				
	Total	Employment		Sex	
		Job	No Job	Male	Female
Total	1,000	341	659	511	489
Agria	293	121	172	145	148
Wheaton	144	60	84	70	74
Greenham	94	38	56	44	50
Newbay	55	23	32	31	24
Induston	707	220	487	366	341
Oakdale	61	26	35	36	25
Crowdon	244	73	171	128	116
Smokeley	147	49	98	80	67
Mudwater	255	72	183	122	133

Source: Samplonian Statistical Office.



Much attention has been paid to the layout of the table, because the tables produced should be camera-ready. Therefore there are many options in Abacus to control the layout. It is possible to specify up to 10 lines of text for the header and for the footer of the table, and one can select both horizontal and vertical rules (as in the example), only horizontal rules or no rules at all. The layout of the text in column headers and the width of the columns can also be influenced.

A rounding procedure can be carried out to protect confidential data in the table. Cell totals, but also marginal totals are rounded to a multiple of some specified constant, *e.g.* 5. Abacus can provide both normal rounding and random rounding. If the user is not satisfied with the resulting table, he can import the output of Abacus into the spreadsheet program Lotus 123, and carry out further processing there. A final feature to be mentioned here is the possibility of creating new variables by recoding existing variables (*e.g.* from age to age classes). More details about Abacus can be found in Bethlehem *et al.* (1989a).

## 9. ANALYSIS

The CBS has not developed any software for statistical data analysis, the main reason being that there are already enough good statistical packages available. The CBS itself uses the packages SPSS (both on mainframe and micro) and Stata (on micro). To make these packages part of the integrated system for survey processing, tools have to be available to export the data from Blaise to them. This is realized in two steps. First, the data file is converted from the Blaise format to Ascii format, and second, the information about the variables, as available in the Blaise questionnaire, is translated in such a way that it can be understood by the particular package. Thus, a setup file is created. By running this setup from within the statistical package, a system file is created. And by loading the system file, the user can start straight away with his analysis, without having to bother about specifying the variables, labels, *etc.*

The procedure above only works for SPSS and Stata, and not for other packages. Of course, this approach could be implemented for every known statistical package, but that would require a large programming effort. Instead, a different road was taken. The Blaise System has a special setup generator utility. The user "paints" the structure of the setup file in a word processor, and by running the setup generator with this general setup description and the Blaise questionnaire as input, a real setup file is created. So, with the setup generator the user can generate setup files for his own favourite package.

## 10. CONCLUSION

The advent of the microcomputer has had a considerable impact on the work of the national statistical office. The subject matter statistician is making use of it more and more, and for his work, he needs an integrated survey processing system like the one based on the Blaise System. The power of this system lies on the one hand in the consistency it enforces in the various steps of data collection and data processing. This makes the whole process easier to manage and to control. On the other hand it also encourages standardization between different departments. Since all departments use the same software for their surveys, exchange of information between departments is easier and less error prone.

The integrated approach to survey processing was developed with in mind a highly centralized organization, like that of the Netherlands Central Bureau of Statistics. In such an organization, this approach can lead to a substantial increase in efficiency. However, not all statistical offices have a centralized structure. Particularly in larger countries, data processing is often

decentralized. Regional offices take care of data processing in their own regions, and the resulting data files are sent to the central office. In the central office, the regional files are combined into one national file. The integrated approach can also be applied successfully in such an environment: the central office develops the Blaise questionnaire, and copies of the generated data entry program are sent to each regional office. This ensures consistency of data collection and data editing at the regional level. The regional data files will all have the same Blaise format, so combining them into one national file will be a simple job using the tools of the Blaise System. Since all regional files will be "clean", no further editing will be necessary at the national level. The only job for the national office will be to tabulate and analyse, and to publish the results. Furthermore, if either the regional offices or the central office can make regional publications.

The Blaise System has been tested and used since the middle of 1986 for a substantial number of surveys. The system is developed in close cooperation with the users. Every new version contains enhanced features. Abacus has been in use for over a year, and is very popular among its users. The Bascula program is still in development. The first prototype has just been released.

We did not mention the possibility of exporting data and meta-information from the Blaise System to the Paradox database package. In the future, a link will also be established between Blaise and the Oracle database system. In this way, a client/server architecture can be realized for Blaise users.

At the end of the statistical production chain, there are some aspects of publication that still have to be dealt with. In the first place, software will be developed to assess the risk of disclosure of confidential (private) information in statistical information to be published. Tools will also be offered to protect tables or data files against these risks. Finally, statistical offices engage more and more in electronic publication of statistical information, *i.e.* statistical information on floppy disks, CD-ROM, *etc.* To help the users of this type of information in selecting the subset of information they need, user-friendly software must be made available to them. This software is now being developed.

## REFERENCES

- BETHLEHEM, J.G., VAN BUITENEN, A.A.A., HUNDEPOOL, A.J., ROESINGH, M.J., and VAN DE WETERING, A. (1989a). Abacus 1.0, A Tabulation Package, Compact Guide. CBS report, Voorburg: Netherlands Central Bureau of Statistics.
- BETHLEHEM, J.G., HUNDEPOOL, A.J., SCHUERHOFF, M.H., and VERMEULEN, L.F.M. (1989b). Blaise 2.0/An Introduction. CBS report, Voorburg: Netherlands Central Bureau of Statistics.
- BETHLEHEM, J.G., HUNDEPOOL, A.J., SCHUERHOFF, M.H., and VERMEULEN, L.F.M. (1989c). Blaise 2.0/Language Reference Manual. CBS report, Voorburg: Netherlands Central Bureau of Statistics.
- BETHLEHEM, J.G., and KELLER, W.J. (1987). Linear Weighting of Sample Survey Data. *Journal of Official Statistics* 3, 141-154.
- KELLER, W.J., BETHLEHEM, J.G., and METZ, K.J. (1990). The impact of micromputers on survey processing at the Netherlands Central Bureau of Statistics. *Proceedings of 1990 Annual Research Conference, U.S. Bureau of the Census*, 637-645.

## Research and Testing of Telephone Survey Methods at Statistics Canada

J. DOUGLAS DREW<sup>1</sup>

### ABSTRACT

Findings from the research and testing of telephone and computer assisted survey methods for household surveys are presented, followed by discussion of how these findings will influence the redesign of household surveys at Statistics Canada during the 1990's. Significant emphasis is given in the presentation to the Canadian Labour Force Survey.

**KEY WORDS:** Data collection; Household surveys; Sample design.

### 1. INTRODUCTION

The 1980's have seen significant changes in survey taking due to advances in technology and the development of modern telephone survey methods, and the pace of these changes will likely accelerate during the 1990's. In this paper we will describe the research, testing, and development of methods that will form the infrastructure underpinning the data collection activities for Statistics Canada's household surveys during the 1990's. This research has focused in particular on the Canadian Labour Force Survey, and was carried out from 1985 to 1989 with a view to identifying improvements to be implemented during the 1991 post-censal redesign of the survey.

### 2. RESEARCH AND TESTING PROGRAM FOR LFS

The Canadian Labour Force Survey (LFS) is the largest household survey conducted by Statistics Canada, with a sample size of 62,300 households per month. It follows a rotating panel design in which households remain in the sample for six consecutive months, after which they are rotated out. It is based on a multi-stage area sample, with a decentralized interviewing staff of 1,000 local interviewers located across Canada and reporting to one of five Regional Offices.

Until the early 1970's, all interviewing was face to face. In 1972 telephone interviewing was introduced in large urban areas for follow-up interviews with households after they had received a face to face interview during their first month in the sample. In the literature, such telephone follow-up is referred to as "warm telephoning", to distinguish it from "cold telephoning" where the telephone interview is not preceded by a face to face interview (Groves *et al.* 1988).

The warm telephoning was initially restricted to major urban areas due to the frequency of party lines in smaller urban and rural areas and concerns this raised about the confidentiality of the data being collected. However, during the 1981 redesign of the survey, warm telephone interviewing was tested for the small urban and rural areas, and it was found respondents were willing to be interviewed by telephone, and the procedure had no impact on response rates or survey estimates (Choudhry 1984). The extension of telephoning to these areas in 1984 resulted in a 10% reduction in the data collection costs for the survey.

---

<sup>1</sup> J. Douglas Drew, Assistant Director, Household Surveys Division, Statistics Canada, Ottawa, Ontario, K1A 0T6.



In 1985, following introduction of a redesigned sample, a research program was started to investigate what further improvements in data collection could result from: (i) more use of the telephone in collection, (ii) telephone survey methods where the telephone is used both for sampling and for data collection, and (iii) Computer Assisted Interviewing (CAI) methods.

In the study of telephone and CAI methods it was useful to characterize the survey design in terms of a number of design factors as follows:

- (i) **Mode of collection.** As already noted, the current mode of collection is warm telephoning, with an initial face to face interview and predominantly telephone interviews in later months. Alternative modes of collection include cold telephone interviewing with face to face follow-up of telephone nonrespondents, and cold telephoning without face to face follow-up.
- (ii) **Organization of interviewers.** The LFS currently features a local organization, with the local interviewers doing a mixture of face to face interviews and telephone interviews from their homes. An alternative is a central organization, with interviewers working out of one or more central sites – in the case of Statistics Canada, its five Regional Offices across the country. A third organizational model is a mixed one where interviewing is done by a combination of local and central interviewers.
- (iii) **Technology.** The current technology for the survey is traditional paper and pencil. The alternative technology considered most viable for household surveys is Computer Assisted Interviewing. CAI is commonly referred to as CATI (Computer Assisted Telephone Interviewing) when done centrally, and CAPI (Computer Assisted Personal Interviewing) when done locally using portable computers that interviewers use for face to face interviews and for telephone interviewing from their homes.
- (iv) **Frame and sample design.** The survey has been based on an area sampling design since its inception in 1945. Alternatives include: telephone frames based on either Random Digit Dialing (RDD) methods (Waksberg 1978) or a combined use of RDD and list frames of published telephone numbers; other list frames which unlike telephone lists, are conceptually complete; and dual frame methods combining two or more of the above frame options.

In the following sections findings from the research and testing program pertaining to each of these design factors are discussed.

### 3. MODE OF COLLECTION

A major test was carried out from 1985 to 1989 to determine the impact of cold telephone interviewing with face to face follow-up as an alternative to the current warm telephoning. This test was referred to as the Telephone First Interview Test. The test was embedded into the ongoing LFS in urban areas of Quebec and Ontario. The methodology is reported fully in Drew, Choudhry and Hunter (1988). In brief, newly sampled LFS dwellings were matched to lists purchased from the telephone companies on the basis of address information. Match rates of 65% were obtained. Test and control samples were selected from the matched dwellings such that each test dwelling was paired with a control dwelling within the same sampling unit (city block). For the test dwellings, telephone numbers were provided to interviewers, who were to attempt a telephone interview, but to use face to face follow-up if necessary. Interviewers were unaware of the existence of a control sample, and followed normal procedures for all dwellings for which telephone numbers were not provided.

Table 1

Telephone First Interview Test (October 1985 – March 1989)  
Estimate for Test Treatment as Percent of Estimate for Control Treatments

Characteristic	Ontario		Quebec	
	Percent	t-statistic	Percent	t-statistic
Employment	98.5	-1.22	97.2	-1.64
Unemployment	96.3	-0.94	111.8	2.27*
Not in LF	101.1	0.88	98.2	-0.63
Pop. 15 +	909.2	-0.86	98.5	-1.31
Pop. in Hhld = 3 +	97.8	-0.76	94.1	-1.26
Pop. in 1 person Hhlds	100.6	0.25	104.1	-0.93
Pop. in 2 person Hhlds	101.1	0.26	101.0	0.43
Pop. in 3 person Hhlds	98.6	-0.56	94.7	-1.28
Emp. Male 15-24	95.2	-0.75	88.7	-2.00*
Emp. Male 25 +	98.5	-1.11	96.7	-1.54
Emp. Female 15-24	99.0	-0.11	111.5	1.44
Emp. Female 25 +	99.2	-0.65	97.1	-1.11
Unemp. Male 15-24	105.6	0.53	119.0	1.53
Unemp. Male 25 +	94.5	-0.80	96.7	-0.31
Unemp. Female 15-24	99.6	-0.15	119.4	1.30
Unemp. Female 25 +	90.9	-1.22	123.9	2.71**
Not in LF Male 15-24	99.8	-0.07	99.9	0.47
Not in LF Male 25 +	105.6	1.57	101.3	0.07
Not in LF Female 15-24	101.9	0.56	95.3	-0.56
Not in LF Female 25 +	99.2	-0.14	97.3	-0.92
Pop. Male 15-24	97.2	-0.49	95.2	-0.73
Pop. Male 25 +	99.9	-0.09	97.7	-1.49
Pop. Female 15-24	99.9	0.16	105.4	0.95
Pop. Female 25 +	98.9	-1.28	98.4	-1.24

\* t-statistic significant at 5% level

\*\* t-statistic significant at 1% level

No significant differences in response rates were found between the test and control samples. For Quebec, response rates were 96.1% for both samples, while for Ontario, the rate of 96.3% for the test sample was marginally lower than that of 96.5% for the control sample.

When comparing the labour force estimates obtained in the test and control samples, certain estimates from Quebec early in the test for the period October 1985 to February 1987 were significantly different. In particular the employed and unemployed males in households of three or more persons were underestimated in the test sample (see Drew, Choudhry and Hunter 1988). Table 1 presents data over the full life of the test from October 1985 to March 1989. For Quebec, a few statistically significant differences existed – which stemmed from the influence of the earlier time period. When the data are analyzed from March 1987 onwards, these differences are not significant. In Ontario, there were no significant differences. Speculating on the differences in Quebec, their co-incidence with a program of inspection of welfare recipients carried out by the provincial government suggests that measures external to the survey led to a climate in which there was decreased trust of cold telephone interviews. We were fortunate

**Table 2**  
Nonresponse Rates: Warm Telephoning versus Cold Telephoning  
with/without Face to Face Follow-up

Method	Test 1	Test 2
Warm telephoning with letter (ongoing LFS)	4.1	5.6
Cold telephoning with letter	6.9	9.8
Cold telephoning without letter	8.5	–

Test 1: October 1985 – September 1986; Ontario and Quebec  
Test 2: July 1988 – March 1989; Nova Scotia and Alberta

to have been conducting the test during this period, because the finding that survey results obtained under cold telephoning are more subject to external influences than are those obtained under warm telephone interviewing will be an important consideration in any decisions on extension of telephone interviewing.

Cold telephoning without face to face follow-up was also studied. Two Telephone Sampling Tests were carried out in which the LFS was conducted as a central telephone survey, with interviewing from the Regional Offices. Nonresponse rates for the two tests and comparable rates for the ongoing LFS are presented in Table 2.

The first test studied two sampling methods – RDD, and a combination of list sampling for published numbers and RDD for nonpublished numbers. The list sampling featured introductory letters, but the RDD sampling did not. Differences in response rates would seem to point to the positive effects on response rates of an advance letter. For both tests, the comparison of warm versus cold telephone interviewing revealed nonresponse rates for cold telephoning which were higher at a 5% significance level. The second test was based solely on a list sample of published numbers.

An important issue is the nonresponse bias, if any, resulting from the higher nonresponse under cold telephoning without face to face follow-up. As a proxy to these extra non-respondents, Laflamme (1990) looked at non-first-month-in-sample households from the ongoing LFS who had a telephone, but who were interviewed face to face. He found that size of the proxy group, at 3.5% of respondents, was close to the size of the extra nonresponse under cold telephoning without face to face follow-up. Further, he found the unemployment rate for the proxy group was 12.8%, versus 7.4% for persons in households interviewed by telephone. Exclusion of the proxy group from the sample would have lowered the national unemployment rate from 8.1% to 7.9%. This is clearly a serious bias, given the accuracy required for the LFS national estimates. As the proxy assumption seems a reasonable one, these findings raise serious concerns about cold telephone interviewing without face to face follow-up for the LFS.

Table 3 compares unemployment and participation rates for the first telephone sampling test with corresponding estimates from telephone households in the LFS. The only estimate found to be significantly different at a five percent level from estimates produced for the telephone population from the ongoing LFS was the unemployment rate for Quebec for the RDD treatment. It is worth noting that the test was carried out at the same time that problems emerged with estimates for Quebec in the Telephone First Interview Test. Another point worth noting is that while other differences in unemployment rates were not statistically significant, the rates for cold telephoning were higher. Other researchers have observed differences in the same direction, also without being able to attribute statistical significance to them. These data might benefit from a meta analysis.



**Table 3**  
Telephone Sampling Test (October 1985 – September 1986)  
Unemployment and Participation Rates

Province	Design	Unemployment Rate (S.D.)		Participation Rate (S.D.)	
Quebec	LIST	12.3	(0.78)	64.1	(1.08)
	RDD	13.0*	(0.88)	62.8	(1.28)
	LFS	10.9	(0.27)	63.4	(0.29)
Ontario	LIST	7.3	(0.59)	69.0	(1.11)
	RDD	7.9	(0.63)	69.0	(1.18)
	LFS	6.9	(0.16)	69.0	(0.20)

\* Significant difference between RDD and LFS Unemployment rates for Quebec

In summary, the test results showed cold telephoning without face to face follow-up yielded higher nonresponse rates than the current warm telephoning method, and while inconclusive, there was some evidence that it yielded higher unemployment rates. On the other hand, cold telephoning with face to face follow-up, apart from the one period of time in Quebec, was found to yield data comparable to that under warm telephoning.

On the basis of these findings, it was decided to implement cold telephoning with face to face follow-up for the LFS apartment frame sample, which constitutes roughly 4% of the overall sample. The availability of the telephone number for apartment frame units, it was reasoned, would help overcome problems in gaining access to highrise apartment buildings, and allow for more attempts to find persons at home than is feasible with face to face interviewing. These expectations seem to have been borne out. As reported by Dufour (1990), while first month nonresponse rates for the apartment sample continue to be higher than corresponding first month rates for the non-apartment sample, the gap has narrowed from a difference of 8.7 percentage points in the year before the change to a difference of 5.7 percentage points during the first five months under the new procedure.

Another change to the mode of collection for the ongoing LFS was to introduce telephone follow-up of the first month in sample households which could not be contacted during an initial visit to the dwelling. This procedure was introduced in 1986, and led to a \$100,000 per year savings in data collection costs.

The combined effect of the telephone first interview for the apartments, and the telephone follow-up for first month nonrespondents has been an increase in the overall telephoning rate for the survey from 80% in 1985 to 83% in 1990.

#### 4. ORGANIZATION OF INTERVIEWING STAFF

During the testing program, two alternatives to the current local organization of the interviewing staff were studied. The telephone sampling tests already described considered a "central" organization where all of the interviewing was done out of the Regional Offices. Another test examined a mixed organization, in which the current warm telephoning mode of collection was retained. The test of the mixed organization was carried out from January 1988 to March 1989 in two Census Metropolitan Areas in which Regional Offices are located – Montreal and Halifax. Its primary objective was to measure the cost implications of such a mixed organization.

The test methodology consisted of face to face collection by local interviewers for first month in sample cases, and telephoning by central interviewing staff working out of the Regional Offices for most non-first month cases. Whenever nonresponse follow-up was required for households initially assigned to the central interviewers, this was carried out by the local interviewers. This methodology was initially tested for the Labour Force Survey by Muirhead *et al.* (1975) and has been extensively studied by the United States Bureau of the Census (1987), where the centralized interviewing is being done using Computer Assisted Telephone Interviewing (CATI).

One of the complexities of the method was the practice followed for the first half of the test of transferring cases requiring nonresponse follow-up from the central to the local interviewers at the mid-point of the interviewing week. For the second half of the test, this so-called re-cycling was restricted to cases where the telephone number was determined to be no longer valid. During the first half of the test, nonresponse rates were 8.0% for the test treatment versus 6.1% for the control procedures corresponding to the decentralized interviewing used for the ongoing LFS. The gap narrowed to 7.3% versus 6.7% during the second half.

From the first telephone sampling test, interviewing costs per household were estimated to be \$2.72 for central data collection with telephone list sampling, versus \$3.53 for RDD sampling. The extra costs for RDD methods is due to the time spent in screening for residential telephone numbers. These costs include \$0.46 per household for long distance charges. This amount was estimated based on long distance rates and data on length of calls, since record keeping practices in the regional offices did not permit the extraction of actual costs incurred. Comparable costs for the ongoing LFS were \$4.76 per household for interviewer fees and expenses. The test of the mixed organization yielded savings relative to the ongoing LFS of \$0.78 per household in interviewer fees and expenses. The above cost comparisons do not factor cost of office space and equipment into the costs under the centralized and mixed organizations. Nor do they consider the costs of transferring documents to and from local interviewers under the mixed and local organizations, which under the current paper and pencil technology is accomplished by express mailing of documents, but under CAI scenarios would be transmitted electronically.

The mixed organization was considered only for Regional Office cities, as extension beyond Regional Office cities would imply greater long distance telephoning. More importantly, the sample design for smaller urban and rural areas is clustered so that primary sampling units yield sample sizes corresponding to an interviewer assignment. Centralization of the telephone portion of the sample would necessitate more clustering of the sample in order to retain a sufficient workload for the local interviewers. Also in medium sized urban centres where there are currently four to five interviewers, the number of local interviewers under the mixed organization would be reduced to one or two, significantly reducing the flexibility to have interviewers fill in for one and other during vacations and illness.

Of the three organizational models considered, all had advantages and disadvantages. The local organization yielded the lowest nonresponse rates, albeit at the highest per unit data collection cost. The mixed organization had marginally higher nonresponse and marginally lower costs, and was limited in where it could be applied. In the final analysis, the mixed organization was seen as introducing a lot of complexity for at best marginal gain. The central organization, which offered substantial savings in data collection costs, resulted in a 68-75% increase in nonresponse relative to the local organization. As discussed in section (3), there is evidence that this extra nonresponse would introduce a serious nonresponse bias into the LFS estimates. Moreover there are concerns that the gap in nonresponse rates attainable under local versus central organizations might widen in the future, as increasing exposure to telephone

solicitation and increasing availability of telephone screening technology renders the population less receptive to telephone interviewing. Such developments favour survey design strategies which, although they may allow for flexibility to do telephoning, also allow for face to face follow-up wherever needed. The local organization best offers this flexibility. On the basis of the above considerations it has been decided to retain the current local organization.

## 5. TECHNOLOGY

Catlin, Ingram, and Hunter (1988) carried out a controlled study comparing CATI and paper and pencil interviewing. In the study the LFS questionnaire was administered to RDD samples of 1,000 households per month per treatment over a period of nine months. All interviewing took place from Statistics Canada headquarters in Ottawa.

The study was part of a collaborative research effort with the United States Bureau of the Census (USBC), and the CATI software used was developed by the USBC. The wording of the questionnaires was purposively the same for both treatments. Features unique to the CATI treatment were automatic branching, some basic on-line edits, and automated call scheduling.

Three quality improvements were discernable for CATI relative to paper and pencil methods. First, the overall rate of edit failures during post-collection data processing was 50% lower for CATI. Second, there was a virtual elimination of branching errors under CATI. Importantly, this occurred for certain portions of the questionnaire, which, although infrequently encountered, have a bearing on determination of labour force status, and which under paper and pencil interviewing are subject to high levels of branching errors. Third, the average household size reported under CATI was 3% higher, which represents roughly a 50% reduction in the underenumeration in the LFS relative to the Census. This improvement seems to stem from the enforced probing built into the CATI instrument for additional household members and for persons temporarily away.

Based on these findings, it has been decided that the introduction of Computer Assisted Interviewing should be one of the major thrusts of the 1991 post-censal redesign of the LFS. Due to the preference for maintaining a local organization of interviewing staff, a CAPI implementation is being planned for.

## 6. FRAME AND SAMPLE DESIGN

### Telephone Frames

Telephone coverage and the extent to which characteristics of those without telephones differ from characteristics of those with telephones are important factors in the design of telephone survey methods – particularly as regards frame strategies.

In an international review of telephone coverage, Trewin and Lee (1988) found telephone coverage in Canada to be one of the highest in the world at 97-98%. As is typical of the situation in most countries they surveyed, persons in non-telephone households in Canada tend to have lower incomes and higher rates of unemployment.

Table 4 gives the percentage of non-telephone households in Canada from 1976 to the present. Telephone coverage, while already high in 1976 has been steadily edging upwards, although it appears to have levelled off over the last few years at around 98.5%.



**Table 4**  
Non-telephone Households by Province (%)

	1976	1981	1985	1987	1990
Canada	3.5	2.4	1.8	1.5	1.5
Newfoundland	10.0	6.0	5.1	3.6	1.9
Prince Edward Island	—	—	—	—	2.8
Nova Scotia	7.5	4.6	3.5	3.2	1.5
New Brunswick	5.8	5.3	5.3	3.3	2.2
Quebec	3.3	2.1	1.6	1.5	1.5
Ontario	2.5	1.9	1.0	1.0	1.2
Manitoba	4.1	2.3	2.7	2.4	1.7
Saskatchewan	3.6	2.5	2.3	2.4	2.3
Alberta	3.0	2.4	2.0	1.8	2.0
British Columbia	4.2	2.8	2.4	1.3	1.5

Source: Statistics Canada, Estimates from Household Facilities & Equipment Survey

**Table 5**  
Labour Force Characteristics by Telephone Status

Province	Telephone Status	Unemployment Rate	Participation Rate
Nova Scotia	published	9.0	71.9
	non-published	9.8	70.2
	non-telephone	17.2	62.3
Alberta	published	6.3	80.7
	non-published	8.2	81.5
	non-telephone	11.1	67.0

Laflamme (1990) undertook a study comparing characteristics of the non-telephone and telephone universes. The study included a breakdown of those with published versus non-published numbers, obtained by linking telephone numbers supplied by LFS respondents to lists of published telephone numbers. Two provinces were included in the study, Nova Scotia and Alberta. For Nova Scotia 9.7% of numbers, and for Alberta 11.2% of numbers were found to be non-published. Unemployment and participation rates reported by Laflamme are given in Table 5.

This study replicates findings from earlier studies that the labour force characteristics of persons without telephones are very different from those with telephones. The labour force characteristics differ but to a lesser extent between persons with published versus non-published numbers.

While the non-telephone population accounts for only 1.0%-1.5% of the population, the differences in labour force characteristics are sufficiently large that simply excluding the non-telephone population is not a viable option for the LFS, given the accuracy required for the national employment and unemployment estimates (coefficients of variation of 0.5% and 2% respectively).

Another difficulty with telephone frames, particularly for panel surveys is their rapid deterioration. Drew, Dick and Switzer (1989) found a 0.5 – 1.0 % rate of additions and deletions to the stock of published residential numbers per month. Hence telephone samples cannot remain representative of the telephone universe for very long unless they are updated. The authors proposed a strategy of updating samples for a panel survey using files of published numbers acquired on an ongoing basis from telephone companies. An operational test of the procedure over a nine month period was a success. Their procedure applied only to published numbers sampled from a list frame, and did not provide a solution to the problem of keeping a sample selected using Random Digit Dialing methods up to date over the life of a panel.

Because of the coverage and updating problems with telephone frames, approaches where dwellings as opposed to telephone numbers are the sampling units are seen as having more promise for large scale panel surveys. It is worth noting that the situation can be quite different for other surveys. Catlin *et al.* (1984) showed that coverage biases for general population characteristics were less than for labour force characteristics. Further for smaller surveys (e.g., those with sample sizes of 10,000 or less) small biases are less important given the larger relative sampling errors for these surveys. These findings led to establishment of an RDD household survey capacity in 1986. It has been used for numerous one-time surveys and for the General Social Survey, which is an annual survey of 10,000 households.

### Area Frame

As has already been described, it is possible in urban areas to match selected addresses from an area frame to telephone lists in order to permit cold telephone interviewing. The experiences with the LFS have been that telephone numbers can be obtained in this fashion for approximately 60% of households. These match rates are based on exact matching after standardization of the address information, and they could be improved through use of record linkage methods. With telephoning for a substantial portion of the first month cases, the clustering of the sample could be reduced somewhat, but a clustered sample remains a constraint imposed by an area frame design.

It is planned to investigate the feasibility of extending these procedures to rural areas, which would entail changing the type of information collected when dwelling lists are created for the survey. The information currently collected tends to be descriptive of the physical characteristics of the dwelling, whereas to successfully match with lists of telephone subscribers, information such as name (often readily available on mail boxes), street name and number, or in their absence the rural route number and postal code, would be required.

### Address Register

Statistics Canada is constructing an Address Register in urban areas of Canada. It will be used in the 1991 Census to improve coverage by providing an independent check on the dwelling lists created by the Census enumerators (Drew, Royce and van Baaren 1989). The Address Register will be a machine readable list of addresses constructed by linkage of various administrative data sources, including lists of customers with published numbers purchased from telephone companies. During the use of the Address Register in the 1991 Census, its coverage will be updated to correspond to that of the 1991 Census.

It is planned during the 1991 post-censal redesign of the Labour Force Survey to conduct studies into use of the Address Register as a frame for household surveys in urban areas. If the conclusion is that it should be adopted as a frame, the Address Register will be updated on an ongoing basis following the 1991 Census.

An advantage of the Address Register as a frame over the area frame is that telephone and non-telephone households are known ahead of time. Hence the two can be sampled as separate strata – with a reduced amount of clustering for the telephone stratum, for which a significant portion of the first month interviews could be done by telephone. The non-telephone stratum would include those households with non-published numbers and those households without telephones. Evidence from earlier studies showed the refusal rate when cold telephoning households with non-published numbers to be 12% compared to 4% for all households. Under warm telephoning there is no corresponding increase in the refusal rates for households with non-published numbers, and there is a good success rate in converting these households to respond by telephone in later months. This finding and the desire to be sensitive to privacy concerns of individuals support the face to face interview of such households, which account for an estimated 10-15% of numbers.

### Dual Frame

In urban areas, if the coverage of the Address Register as the sole frame is not adequate, a dual frame design in which the Address Register is supplemented by a small area sample will be considered. There are different forms the supplementary sample could take. A promising option, not involving the expense of building and maintaining both a conventional area frame and an Address Register, would be to use an interval approach in which a sample of consecutive dwellings on the Address Register would be selected and checked in the field. Any dwellings found between the Address Register dwellings constitute a sample of dwellings missing from the Address Register.

Mian (1990) has studied dual frame methods considering a cost and variance optimization for the general case where neither of the frames needs to cover the entire universe. This was felt to be a practical model since the area frame, while conceptually complete, in practice suffers from 3-4% undercoverage relative to the Census, in addition to the 5% of the population which is not represented because of nonresponse. Extension of Mian's model to include a non-sampling error component will permit factoring into the optimization what we know or may wish to assume about the coverage and nonresponse biases under alternative frame approaches. It can be used in the context of dual frames combining the Address Register and area frames in urban areas, and combining area and telephone frames in rural areas.

## 7. 1991 POST-CENSAL REDESIGN OF LFS

The Labour Force Survey is redesigned following each decennial population census. Redesigns have normally focused on redesign of the sample, but in the 1970's a major revision was carried out encompassing a sample redesign, changes to the questionnaire content, wording of questions, and survey outputs, and a major overhaul of the survey processing systems including introduction of a network of mini-computers in the regional offices to support survey operations and regional data capture. In contrast, redesign efforts during the 1980's were restricted to a sample redesign.

While decisions on the scope of the 1991 post-censal redesign have yet to be taken, an effort falling somewhere between the major revision in the 70's and minimal redesign in the 80's appears needed. The work on the redesign is at an early planning stage, and is proceeding through four sub-projects focusing on: (i) content and questionnaire issues, (ii) modernization of the survey processing systems and review of survey outputs, (iii) development, testing, and implementation of Computer Assisted Interviewing, and (iv) sample redesign. (Drew *et al.* 1991).



Sub-projects (iii) and (iv) are those which will be concerned with telephone and CAI methods. Current plans for these sub-projects as they relate to the survey design factors described in earlier sections of this paper are briefly summarized below.

### **Technology and Organization**

Based on the positive findings from testing of Computer Assisted Interviewing for the LFS reported by Catlin *et al.* (1988), it has been decided to make the adoption of CAI one of the major thrusts of the redesign. Moreover, for reasons already discussed, a decision has been taken that the current local organization of the interviewing staff should be retained, so that the implementation of CAI methods will take the form of Computer Assisted Personal Interviewing (CAPI). Specifically, local interviewers will be equipped with lightweight portable computers that they will carry with them for face to face interviewing, and that they will use to conduct telephone interviews from their homes. The mix of face to face and telephone interviewing may remain much the same as it is currently – 83% telephone and 17% face to face – or it may shift to more telephoning if it is decided to adopt more cold telephone interviewing of households during their first month in the sample.

The work plan for the Computer Assisted Interviewing sub-project includes a field test during 1991 of a touch screen portable computer, and in later years development or acquisition of CAI software, a combined test of CAI and questionnaire alternatives, development of on-line editing, and an automated version of the interviewer manual embedded into a help screen accessible during interviewing.

### **Frame and Mode of Collection**

A key research finding was that cold telephone interviewing with face to face follow-up yields response rates and labour force estimates comparable to those under the current warm telephone collection procedure for the LFS, which features face to face interviewing for the first month households in the sample. One exception was observed in Quebec, where, as discussed, for a period of time data differences were seen. The importance of face to face follow-up in maintaining response rates favours the retention of the dwelling as the sampling unit, and supplying the telephone number to interviewers to give them greater flexibility to use both telephone and face to face interviewing in obtaining first month interviews.

In urban areas, both the current area frame and the Address Register as a frame are consistent with this approach. In rural areas, as described earlier in the paper, research into the feasibility of matching area frame addresses to telephone lists to provide interviewers with telephone numbers will be studied. It is also planned to continue to investigate dual frame methods which in urban areas might consist of the Address Register and an area frame, and in rural areas an area frame and a telephone frame.

## **8. SUMMARY**

The current data collection methodology for the Labour Force Survey consists of a face to face interview for households during their first month in the sample and predominantly telephone interviewing in later months. The introduction of telephone interviewing for the later months took place during the 1970's for major urban areas and during the 1980's for remaining areas. In both instances the introduction of telephone interviewing resulted in significant cost savings without any impact on the response rates or survey estimates. Prior to the telephone and CAI research and testing program begun in 1985, 80% of LFS interviews were done by

telephone. This has moderately increased to 83% through the introduction of telephone follow-up for households which could not be contacted during an initial face to face visit, and by supplying interviewers with telephone numbers for the apartment sample.

The primary benefit of the research and testing program has been to identify the frame and data collection options to pursue during the 1991 post-censal redesign of the survey, including the retention of the current local organization of interviewers, the adoption of Computer Assisted Personal Interviewing, the retention of frame and sample design approaches in which the dwelling is the unit of selection, and the provision of interviewers with telephone numbers to permit the flexibility to use a combination of telephone and face to face interviewing to obtain first month interviews.

## REFERENCES

- CATLIN, G., CHOUDHRY, H., and HOFMANN, H. (1984). Telephone ownership in Canada. Internal report, Statistics Canada.
- CATLIN, G., and INGRAM, S. (1988). The effects of CATI on cost and data quality. *Telephone Survey Methodology*, (Eds. R. Groves *et al.*), 437-450. New York: Wiley.
- CHOUDHRY, G.H. (1984). Results from telephone interviewing experiment in the non self representing areas of the Labour Force Survey. Internal report, Statistics Canada.
- DREW, J.D., and GAMBINO, J. (1991). Plans for the 1991 post censal redesign of the Canadian Labour Force Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, to appear.
- DREW, J.D., CHOUDHRY, H., and HUNTER, L. (1988). Nonresponse issues in government telephone surveys. *Telephone Survey Methodology*, (Eds. R. Groves *et al.*), 233-246. New York: Wiley.
- DREW, J.D., DICK, P., and SWITZER, K. (1989). Development and testing of telephone survey methods for household surveys at Statistics Canada. *Proceedings of the Section on Survey Research Methods, American Statistical Association*.
- DREW, J.D., and JAWORSKI, R. (1986). Telephone survey development on the Canadian Labour Force Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*.
- DREW, J.D., ROYCE, D., and van BAAREN, A. (1989). Address register research at Statistics Canada. *Proceedings of the Section on Survey Research Methods, American Statistical Association*.
- DUFOUR, J. (1990). Implantation de la première entrevue par téléphone pour la base d'appartements de l'enquête sur la population active. Internal report, Social Survey Methods Division, Statistics Canada.
- GROVES, R.M., BIEMER, P.P., LYBERG, L.E., MASSEY, J.T., NICHOLLS, W.L., II, and WAKSBERG, J. (Eds.) (1988). *Telephone Survey Methodology*. New York: Wiley.
- LAFLAMME, F. (1990). Étude comparative entre trois différentes populations visées par l'EPA selon leur type de service téléphonique. Internal report, Social Survey Methods Division, Statistics Canada.
- MIAN, I.U.H. (1990). Dual frame estimation of proportions in sample surveys. Internal report, Social Survey Methods Division, Statistics Canada.
- MUIRHEAD, R.C., GOWER, A.R., and NEWTON, F.T. (1975). The telephone experiment in the Canadian Labour Force Survey. *Survey Methodology*, 1, 158-180.
- TREWIN, D., and LEE, H. (1988). International comparisons of telephone coverage. *Telephone Survey Methodology*, (Eds. R. Groves *et al.*). New York: Wiley, 9-24.
- WAKSBERG, J. (1978). Sampling methods for random digit dialing. *Journal of the American Statistical Association*, 73, 576-579.

# Marginal and Approximate Conditional Likelihoods for Sampling on Successive Occasions

D.R. BELLHOUSE<sup>1</sup>

## ABSTRACT

Marginal and approximate conditional likelihoods are given for the correlation parameters in a normal linear regression model with correlated errors. This general likelihood approach is applied to obtain marginal and approximate conditional likelihoods for the correlation parameters in sampling on successive occasions under both simple random sampling on each occasion and more complex surveys.

KEY WORDS: Likelihood inference; Sampling in time; ARMA models; State space models.

## 1. INTRODUCTION

Consider a finite population of  $N$  units which may be sampled on  $k$  occasions. Let  $y_{ij}$  denote the measurement on the  $j^{\text{th}}$  population unit taken on the  $t^{\text{th}}$  occasion;  $j = 1, \dots, N$  and  $t = 1, \dots, k$ . It is assumed that any two units, say  $j$  and  $j'$ , are independent, but that measurements of the same unit across time are correlated. In particular, assume that for any  $j$ ,

$$(y_{1j}, y_{2j}, \dots, y_{kj})^T \sim N(\mu, \sigma^2 \Omega), \quad (1)$$

where  $\Omega$  is a  $k \times k$  correlation matrix and where  $\mu$  is the  $1 \times k$  vector of fixed means  $(\mu_1, \mu_2, \dots, \mu_k)^T$ . In view of the explicit model assumption in (1), a model-based approach to survey estimation is used in this paper. Based on samples taken over the  $k$  occasions, it is of interest to estimate  $(\mu_1, \mu_2, \dots, \mu_k)^T$ . The form of the model-based estimates  $(\hat{\mu}_1, \hat{\mu}_2, \dots, \hat{\mu}_k)^T$ , if obtained by maximum likelihood or generalized least squares, for example, will depend on  $\sigma^2$  and the parameters in  $\Omega$ . It is therefore necessary to obtain good estimates of  $\sigma^2$  and the parameters in  $\Omega$ .

The notation of Bellhouse (1989) is used to describe the sampling scheme considered here, namely one-level rotation sampling. On any occasion,  $c$  rotation groups are sampled. Rotation group  $r$  ( $r = 1, 2, \dots, k + c - 1$ ), denoted by  $G_r$ , consists of  $m_r$  sample units. On occasion  $t$  ( $t = 1, \dots, k$ ), the sample consists of the units in  $G_t, G_{t+1}, \dots, G_{t+c-1}$ , so that the total sample size on occasion  $t$ ,  $n_t = m_t + m_{t+1} + \dots + m_{t+c-1}$ . On occasion  $t + 1$ ,  $G_t$  is dropped from the sample and  $G_{t+c}$  is added. Each rotation group is chosen without replacement from previously unchosen units in the population. The total sample size over all  $k$  occasions is  $m = n_1 + n_2 + \dots + n_k$ . The maximum number of occasions that a unit remains in the sample is  $c$ .

If  $c$  is small, then estimates of the correlation parameters in  $\Omega$  can be unstable, leading to instability in the estimates of interest  $(\hat{\mu}_1, \hat{\mu}_2, \dots, \hat{\mu}_k)^T$ . Viewed another way, the total number of parameters is at least  $k + 2$  and increases with time, *i.e.* with the addition of new occasions. Since the dimension of the parameter space increases with time, maximum likelihood estimates of parameters may be biased and inconsistent. The problem of the stability of estimates has been addressed in sampling on successive occasions, for example, by Blight

<sup>1</sup> D.R. Bellhouse, Department of Statistical and Actuarial Sciences, University of Western Ontario, London, Ontario, Canada N6A 5B9.



and Scott (1973), who assume that the elements of  $(\mu_1, \mu_2, \dots, \mu_k)^T$  follow a time series process. On using this assumption the dimension of the parameter space is fixed at a relatively small number so that the problems of instability, bias and inconsistency are resolved. In this paper, a different approach is taken. Here the fixed means assumption is retained and marginal and approximate conditional likelihoods are derived for the parameters in  $\Omega$ , treating the fixed means as nuisance parameters.

Marginal likelihood estimation was introduced as a general method for eliminating nuisance parameters from the likelihood function (Fraser 1967; Kalbfleisch and Sprott 1970). Cox and Reid (1987) introduced approximate conditional likelihoods which also address this problem. They argued that the approximate conditional likelihood was preferable to the profile likelihood obtained by replacing the nuisance parameters in the likelihood by their maximum likelihood estimates when the parameters of interest are given. Bellhouse (1990) established the equivalence of marginal and approximate conditional likelihoods for correlation parameters under a normal model.

Following on the work of Cox and Reid, Cruddas *et al.* (1989) obtained an approximate conditional likelihood for the correlation parameter in several short series of autoregressive processes of order one with common variance and autocorrelation parameters. Based on a simulation study, Cruddas *et al.* (1989) showed that the estimate based on the approximate conditional likelihood had a much smaller bias and better coverage properties of the confidence interval than the maximum likelihood estimate from the profile likelihood. The situation described by Cruddas *et al.* (1989) applies directly to sampling on successive occasions in sample surveys. In order to reduce the response burden, individuals in a survey are retained in the sample for relatively short periods of time. It is expected that the use of marginal and approximate conditional likelihoods will improve the estimates of correlation parameters and consequently improve the estimates of the mean for each occasion.

Within a rotation group, the sample measurements on an individual are usually modelled by an autoregressive moving average process (ARMA), *i.e.* the parameters in  $\Omega$  are comprised of the correlation parameters in the ARMA process. See Binder and Hidioglou (1988) for a review of the application of time series models to sampling on successive occasions. Consequently, it is of interest to obtain marginal and approximate conditional likelihoods under ARMA models with application to rotation sampling. The marginal and approximate conditional likelihoods for the correlation parameters in a normal model are obtained in Section 2. The general results of Section 2 are illustrated in Section 3 by applying the results to sampling on successive occasions assuming simple random sampling of units in rotation groups. In Section 4, some methods are given to apply these likelihood methods to complex surveys.

## 2. MARGINAL AND APPROXIMATE CONDITIONAL LIKELIHOODS FOR CORRELATION PARAMETERS UNDER A NORMAL MODEL

Let  $y$  be a vector of sampled observations of dimension  $m \times 1$  which follows the linear model

$$y = X\beta + \epsilon \quad (2)$$

with error vector  $\epsilon \sim N(0, \sigma^2 \Phi)$ , where  $\Phi$  is the  $m \times m$  correlation matrix and where  $\beta$  is the  $p \times 1$  vector of regression coefficients so that  $X$  is  $m \times p$ . The log-likelihood for  $\beta$ ,  $\sigma^2$  and  $\Phi$  is given by

$$L(\beta, \sigma^2, \Phi) = - \left\{ m \ln \sigma + (\ln |\Phi|) / 2 + (y - X\beta)^T \Phi^{-1} (y - X\beta) / (2\sigma^2) \right\}. \quad (3)$$

For a given value of  $\Phi$ ,

$$\hat{\beta} = (X^T \Phi^{-1} X) X^T \Phi^{-1} y$$

and

$$s^2 = y^T \Phi^{-1} y - y^T \Phi^{-1} X (X^T \Phi^{-1} X)^{-1} X^T \Phi^{-1} y \tag{4}$$

are jointly sufficient for  $\beta$  and  $\sigma^2$ .

A marginal likelihood for  $\Phi$  is obtained by making a transformation of the data  $y$  to the sufficient statistics  $\hat{\beta}$  and  $s^2$  and the ancillary statistic

$$a = \Phi^{-1/2} (y - X (X^T \Phi^{-1} X)^{-1} X^T \Phi^{-1} y) / s,$$

where  $\Phi^{-1/2}$  is the  $m \times m$  dimensional matrix such that  $\Phi^{-1} = \Phi^{-1/2} \Phi^{-1/2}$ . The marginal likelihood for  $\Phi$  is the marginal distribution of the ancillary  $a$  times the product of the differentials  $da_i$ ,  $i = 1, \dots, m$ . See Kalbfleisch and Sprott (1970, eqs. 6 and 10) for a general discussion and a general expression for  $\Pi da_i$ . Bellhouse (1978) and, later independently Tunnicliffe Wilson (1989), showed that the marginal likelihood for  $\Phi$  under the normal model is given by

$$L_M(\Phi) = \{ | \Phi |^{1/2} | X^T \Phi^{-1} X |^{1/2} s^{m-p} \}^{-1}. \tag{5}$$

Note that (4) is proportional to the maximum likelihood estimate of  $\sigma^2$  given  $\Phi$  and that  $s^2 (X^T \Phi^{-1} X)^{-1}$  is proportional to the estimated variance-covariance matrix of the maximum likelihood estimate of  $\beta$  given  $\Phi$ . Then (5) can be written as

$$L_M(\Phi) = \frac{ | \text{est var}(\hat{\beta}) |^{1/2} }{ s^m | \Phi |^{1/2} }. \tag{6}$$

To obtain an approximate conditional likelihood, it is first necessary to transform the parameters to achieve parameter orthogonality between the parameters of interest and the nuisance parameters, which now may depend on the parameters of interest. Sets of parameters are orthogonal if the associated information matrix is block diagonal, with each block as the information matrix for each parameter set. The conditional likelihood is related to the distribution of the data  $y$  conditional on the maximum likelihood estimate of the nuisance parameters for fixed values of the parameters of interest. The approximate conditional likelihood is obtained by applying two approximations to this conditional distribution. See Cox and Reid (1987, Section 4.1) for a discussion of the derivation. For example, let  $\Theta$  be the vector of parameters of interest and let  $\Lambda$ , possibly depending on  $\Theta$ , be the vector of nuisance parameters orthogonal to  $\Theta$ . The full likelihood of the data for parameters  $\Theta$  and  $\Lambda$  is denoted by  $L(\Theta, \Lambda)$  and the profile likelihood for  $\Theta$ ,  $L(\Theta, \hat{\Lambda})$  is the likelihood with  $\Lambda$  replaced by its maximum likelihood estimate. The approximate conditional likelihood for  $\Theta$  is

$$L(\Theta, \hat{\Lambda}) | I(\Theta, \hat{\Lambda}) |^{1/2},$$

where  $I(\Theta, \hat{\Lambda})$  is the observed information matrix for  $\Lambda$  at a fixed value of  $\Theta$ . See Cox and Reid (1987, eq. 10).

Following Cruddas *et al.* (1989), Bellhouse (1990) suggested, for model (2), the parameter transformation  $\lambda = \ln \sigma + (\ln |\Omega|)/(2m)$  leaving  $\beta$  the same. The log-likelihood under the new parameterization is denoted by  $L(\beta, \lambda, \Phi)$  and can be obtained from (3). If the entries of  $\Phi$  are functions of a parameter  $\phi$ , then the nuisance parameters  $\lambda$  and  $\beta$  are each orthogonal to  $\Phi$ , *i.e.*

$$-\frac{1}{m} E \left[ \frac{\partial^2 L(\beta, \lambda, \Phi)}{\partial \phi \partial \lambda} \right] = 0$$

and

$$-\frac{1}{m} E \left[ \frac{\partial^2 L(\beta, \lambda, \Phi)}{\partial \phi \partial \beta} \right] = 0,$$

when each entry of  $\Phi$  is a continuous and differentiable function of  $\phi$ . Moreover, in this case the approximate conditional likelihood for  $\Phi$ ,  $L_C(\Phi)$  is the same as the marginal likelihood  $L_M(\Phi)$ , given by (5) or (6). See Bellhouse (1990) for details.

The marginal and approximate conditional likelihood in (5) or (6) can be evaluated at any  $\Phi$  using state space models in the approach of Harvey and Phillips (1979). For any given  $\Phi$ , once the recursions to estimate  $\beta$  and  $\sigma^2$  are complete, the value of  $s^2$  and  $|\Phi|^{1/2}$  can be calculated from Harvey and Phillips (1979, eqs. 5.6 and 6.6, and 4.3 respectively). It is then necessary only to obtain  $X^T \Phi^{-1} X$  and its determinant. The value of  $X^T \Phi^{-1} X$  may be obtained from the final step in the recursive equations of Harvey and Phillips (1979, eq. 3.4).

### 3. SIMPLE RANDOM SAMPLING ON SUCCESSIVE OCCASIONS

#### 3.1 Some General Results for Rotation Sampling

Suppose rotation group  $G_r$  first appears in the sample on occasion  $u$  and last appears on occasion  $v$ . Then  $u$  is either 1 or  $r$  and  $v$  is either  $r + c - 1$  or  $k$ . The total number of occasions on which a unit in  $G_r$  is present in the sample is  $b = v + 1 - u$ . Let  $\bar{y}_{u,r}, \dots, \bar{y}_{v,r}$  be the sample means or elementary estimates for  $G_r$  on occasions  $u, u + 1, \dots, v - 1, v$  respectively. Then under model (1), the contribution of  $G_r$  to the log likelihood in (3) is

$$- \{ bn_r \ln \sigma + (n_r/2) \ln(|\Omega_r|) + [n_r \mathbf{x}_r^T \Omega_r^{-1} \mathbf{x}_r + (n_r - 1) \text{tr}(\Omega_r^{-1} S_r)] / (2\sigma^2) \}, \quad (7)$$

where  $\mathbf{x}_r^T$  is the  $1 \times b$  vector  $(\bar{y}_{u,r} - \mu_u, \bar{y}_{u+1,r} - \mu_{u+1}, \dots, \bar{y}_{v-1,r} - \mu_{v-1}, \bar{y}_{v,r} - \mu_v)$ , where  $S_r$  is the  $b \times b$  matrix of sample variances and covariances for observations within the rotation group, and where  $\Omega_r$  is the  $b \times b$  correlation matrix on the observations on a single unit within the rotation group. Note that the parameters in  $\Omega$  as given in expression (1) will also be the parameters in  $\Omega_r$ . The correlation matrix  $\Omega$  is based on measurements from all occasions 1 through  $k$ ; the correlation matrix  $\Omega_r$  is from the subset of the data observed from occasions  $u$  through  $v$ . By the independence assumption, the full log likelihood is obtained by summing (7) over all rotation groups.

Given the parameters in  $\Omega$ , or equivalently the parameters in  $\Omega_1, \dots, \Omega_{k+c-1}$ , expressions for the maximum likelihood estimates  $\hat{\mu}$  and  $\hat{\sigma}^2$ , for  $\mu$  and  $\sigma^2$  respectively, may be found. Likewise,  $V(\hat{\mu})$ , the estimated variance-covariance matrix of  $\hat{\mu}$  may be obtained. This is illustrated for a first-order autoregressive process in Section 3.2. Then the marginal likelihood for the parameters in  $\Omega_1, \dots, \Omega_{k+c-1}$  is given by (5) with the expressions in (5) given by



$$|\Phi|^{1/2} = \prod_{r=1}^{k+c-1} \Omega_r,$$

$$|X^T \Phi^{-1} X|^{1/2} = V(\hat{\mu})/s^k, \quad (8)$$

$$s^2 = \sum_{r=1}^{k+c-1} \{n_r \hat{x}_r^T \Omega_r^{-1} \hat{x}_r + (n_r - 1) \text{tr}(\Omega_r^{-1} S_r)\},$$

and  $p = k$ , where  $\hat{x}_r$  is  $x_r$  with the  $\mu$ 's in  $x_r$  replaced by their maximum likelihood estimates.

### 3.2 First-Order Autoregressive Processes

When specific forms of the correlation matrices  $\Omega_1, \dots, \Omega_{k+c-1}$  are used, some simplifications to the general form of the marginal likelihood for correlation parameters, given by (5) and (6), may be obtained. For example, assume the first-order autoregressive model

$$y_{tj} = \mu_t + \phi (y_{t-1,j} - \mu_{t-1}) + \epsilon_{tj}, \quad (9)$$

where  $\epsilon_{tj} \sim N(0, \sigma^2)$  for  $t = 1, \dots, k$  and  $j = 1, \dots, N$ , and where the  $\epsilon$ 's are mutually independent. Model (9), essentially Patterson's (1950) model, is a special case of (1). As in Section 3.1, the vector of regression parameters  $\beta = (\mu_1, \dots, \mu_k)^T$ . When the data vector  $y$  contains the measurements on each unit grouped by all the occasions on which it was sampled, as in the rotation sampling description of Section 3.1, the correlation matrix  $\Phi$ , now a function of the autoregressive parameter  $\phi$ , can be written as a direct sum of matrices, each of which are the correlation matrices of a first-order autoregressive process.

The following notation, similar to Patterson (1950), is used to denote various sample sizes, means and sums of squares and cross products (corrected for the appropriate mean) for occasion  $t$ :

- $\pi_t$  = the proportion of units on occasion  $t$  that are matched with units from the previous occasion ( $t - 1$ );
- $n_t$  = the number of units sampled on occasion  $t$ ;
- $\bar{y}'_t$  = the mean of the units on occasion  $t$  that are matched with units from the previous occasion ( $t - 1$ );
- $\bar{y}''_t$  = the mean of the units on occasion  $t$  that are unmatched with units from the previous occasion ( $t - 1$ );
- $\bar{y}_t$  = the mean of all the units on occasion  $t$ ;
- $\bar{x}'_t$  = the mean of the units on occasion  $t$  that are matched with units from the following occasion ( $t + 1$ );
- $syy'_t$  = the sum of squares among units on occasion  $t$  which are matched with units from the previous occasion ( $t - 1$ );
- $syy''_t$  = the sum of squares among units on occasion  $t$  which are unmatched with units from the previous occasion ( $t - 1$ );
- $sxx'_t$  = the sum of squares among units on occasion  $t$  which are matched with units from the following occasion ( $t + 1$ );
- $syy_t$  = the sum of squares among all the units on occasion  $t$ ;
- $sxy'_t$  = the sum of cross products for measurements on sample units from occasion  $t$  matched with sample units from ( $t - 1$ ).

Under the special case of model (9), and after much algebra, it may be shown that (7) summed over all rotation groups  $r$ , the log-likelihood for the data reduces to

$$L(\mu_1, \dots, \mu_k, \sigma^2, \phi) = -m \ln \sigma + (d/2) \ln(1 - \phi^2) - \{A(\mu, \phi) + B(\phi)\} / (2\sigma^2), \quad (10)$$

where  $d$  is the distinct number of units sampled (irrespective of the number of occasions on which a unit is sampled) and  $m$  is the total sample size ( $n_1 + \dots + n_k$ ). Further in (10),

$$A(\mu, \phi) = (1 - \phi^2)n_1(\bar{y}_1 - \mu_1)^2 + \sum_{t=2}^k [\pi_t n_t \{\bar{y}'_t - \mu_t - \phi(\bar{x}'_{t-1} - \mu_{t-1})\}^2 + (1 - \pi_t)n_t(1 - \phi^2)(\bar{y}''_t - \mu_t)^2] \quad (11)$$

and

$$B(\phi) = (1 - \phi^2)syy_1 + \sum_{t=2}^k \{\phi^2 sxx'_{t-1} - 2\phi sxy'_t + syy'_t + (1 - \phi^2)syy''_t\}. \quad (12)$$

For any given value of  $\phi$  the maximum likelihood estimator is  $\hat{\mu} = G^{-1}z$  and  $\hat{\sigma}^2 = \{A(\hat{\mu}, \phi) + B(\phi)\}/m$ , where  $A(\hat{\mu}, \phi)$  is (11) with  $\mu$  replaced with its maximum likelihood estimate and where  $G$  is a symmetric  $k \times k$  band matrix of band width 3 and  $z$  is a  $k \times 1$  vector. The nonzero entries of  $G$  are

$$g_{tt} = \pi_t n_t + (1 - \pi_t)n_t(1 - \phi^2) + \pi_{t+1}n_{t+1}\phi^2, \quad \text{for } t = 1, \dots, k$$

and

$$g_{t,t+1} = -\pi_{t+1}n_{t+1}\phi, \quad \text{for } t = 1, \dots, k-1,$$

where  $\pi_1 = \pi_{k+1} = 0$ . The entries of  $z$  are

$$z_t = \pi_t n_t(\bar{y}'_t - \phi \bar{x}'_{t-1}) + (1 - \pi_t)n_t \bar{y}''_t(1 - \phi^2) - \pi_{t+1}n_{t+1}(\bar{y}'_{t+1} - \phi \bar{x}'_t),$$

for  $t = 1, \dots, k$ , where  $\pi_1 = \pi_{k+1} = 0$  and  $\bar{y}''_1 = \bar{y}_1$ . The vector of estimated means  $\hat{\mu}$  is unbiased for  $\mu$  under model (9) and its variance-covariance matrix is  $\sigma^2 G^{-1}$ . It follows from (5) or (6) that the marginal and approximate conditional likelihood for  $\phi$  is

$$L_M(\phi) = \frac{(1 - \phi^2)^{d/2}}{\{A(\hat{\mu}, \phi) + B(\phi)\}^{(m-k)/2} |G|^{1/2}}. \quad (13)$$

### 3.3 Example

The data for this example are forestry data taken from Cunia and Chevrou (1969, p. 220). The data are the merchantable volume of timber per plot measured on three occasions with partial replacement of the sample units. In rotation sampling it is assumed that once a unit

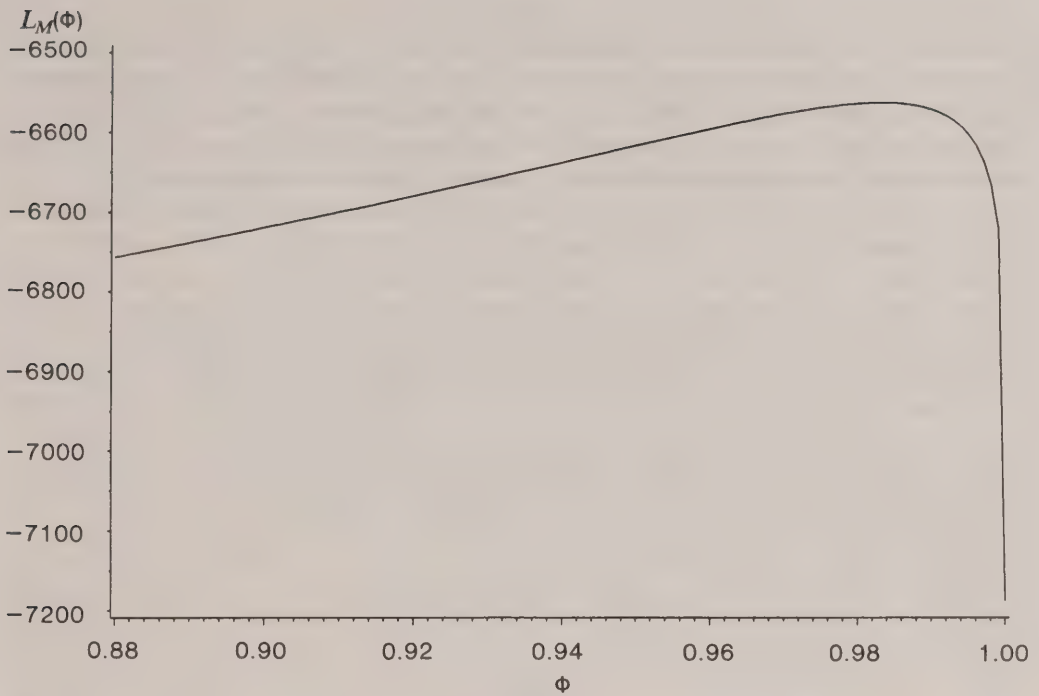


Figure 1. Marginal Likelihood for the AR(1) Parameter

is dropped from the sample it is not selected again. In view of this assumption an adjustment to the data in Cunia and Chevrou was made. In particular, the measurements from sample units matched on the first and third occasions without matching units on the second occasion were dropped from the current example. From the remaining data the following calculations may be made:

$$\begin{aligned} \pi_2 &= 86/139, \quad \pi_3 = 76/100, \quad n_1 = 104, \quad n_2 = 139, \quad n_3 = 100, \quad \bar{y}'_2 = 161.5581, \\ \bar{y}'_3 &= 179.9211, \quad \bar{y}''_1 = 154.0673, \quad \bar{y}''_2 = 167.2075, \quad \hat{y}''_3 = 181.125, \quad \bar{x}'_1 = 147.6512, \\ \bar{x}'_2 &= 163.4342, \quad syy'_2 = 864129.2, \quad syy'_3 = 555369.5, \quad syy''_1 = 943948.5, \quad syy''_2 = \\ &266820.7, \quad syy''_3 = 271762.6, \quad sxx'_1 = 800753.5, \quad sxx'_2 = 559850.7, \quad sxy'_2 = 812435.7, \\ sxy'_3 &= 550943.6, \quad d = 181, \quad \text{and } m - k = 340. \end{aligned}$$

On substituting these data into (13) the marginal and approximate conditional likelihood of the data for the autoregressive order one parameter  $\phi$  may be obtained. This is shown in Figure 1.

#### 4. COMPLEX SURVEYS

There are several ways in which one may proceed to analyze time series data from complex surveys. Each method that can be put forward will depend upon the sample information that is available.



If data are available at the micro level, then variance-covariance matrices based on the complex design can be computed for the elementary estimates for each rotation group. A pseudo marginal likelihood is obtained by replacing  $\hat{x}_r$  and  $S_r$  in (5) and (8) by their complex survey counterparts. A similar approach is taken, for example, by Roberts, Rao and Kumar (1987) in logistic regression analysis for complex surveys: obtain a likelihood or a set of likelihood equations and replace the usual statistics by their complex survey counterparts.

Under simple random sampling,  $S_r$  estimates the finite population variance-covariance matrix for measurements on the occasions covered by rotation group  $r$ . Consequently, in a complex design,  $S_r$  is replaced by a design-consistent estimate of the corresponding finite population variance-covariance matrix. For example, Kilpatrick (1981) looked at a stratified sampling design on two occasions for evaluation of the standing volume of state forests in Northern Ireland; the strata were based on the times, beginning in the 1920's, at which the forests were planted. In order to calculate the stratified sampling equivalent to  $S_r$ , it is necessary to have the estimates of the means on each occasion, strata means, strata variances, and strata covariances for the unmatched and matched samples from the two occasions. For a stratified population, the finite population variance (or covariance) may be decomposed into terms comprising the variation (or covariation) between strata and the variation (or covariation) within strata; see, for example, Cochran (1977, eq. 5.32). Estimates of the means and strata means would be used to obtain a consistent estimates of the between strata variation or covariation component and estimates of the strata variances and covariances would be used to obtain estimates of the within strata variation and covariation. Unfortunately, only certain strata variance and covariance estimates were relevant to Kilpatrick's study, so that there is insufficient published data in the article to calculate a maximum marginal likelihood estimate for the correlation between timber volumes on the two occasions.

In many cases the data at the micro level will not be available. The estimation procedure then depends upon the data that are available. One scenario is considered here; others could be formulated. Suppose that only the elementary estimates and their design effects are available. Let  $\bar{y}_{t,r}$  be the estimate from rotation group  $G_r$  on occasion  $t$  based on a sample of size  $m_r$ . Let  $\text{deff}_{t,r}$  be the design effect associated with  $\bar{y}_{t,r}$ . If  $\sigma^2/m_r$  is the variance of  $\bar{y}_{t,r}$  under simple random sampling, then on appealing to the Central Limit Theorem,

$$(\bar{y}_{t,r} - \mu_t) / (\text{deff}_{t,r})^{1/2} \sim N(0, \sigma^2/m_r) \quad (14)$$

approximately. The modelling may proceed by assuming, within  $G_r$ , an ARMA-type process such as

$$(\bar{y}_{t,r} - \mu_t) / (\text{deff}_{t,r})^{1/2} = \phi (\bar{y}_{t-1,r} - \mu_{t-1}) / (\text{deff}_{t-1,r})^{1/2} + \epsilon_t, \quad (15)$$

where  $\epsilon_t$  has constant variance. This may be easily cast into the framework of model (2), where the data vector  $y$  contains data of the form  $\bar{y}_{t,r} / (\text{deff}_{t,r})^{1/2}$ , where  $\beta$  is  $(\mu_1, \mu_2, \dots, \mu_k)^T$ , and where  $X$  contains entries of the form  $1/(\text{deff}_{t-1,r})^{1/2}$ . The marginal likelihood, obtained as a special case of (5) or (6), may be evaluated using the state space models of Harvey and Phillips (1979) as noted in Section 2. Marginal and approximate conditional likelihood estimation is especially desirable under the model given by (14) and (15). The estimate of  $\phi$  in this case is based on the variation between elementary estimates within each rotation group; the variation within elementary estimates is not available. The length of time a rotation group remains in the sample is short so that the problems of bias and inconsistency in the maximum likelihood estimates will be applicable here.

## 5. DISCUSSION

Binder and Dick (1990) have also suggested the use of marginal likelihood estimation techniques for sampling on successive occasions. In their framework, suppose that the survey estimates of the means, say  $\bar{y}_t$ , are available for each occasion  $t = 1, \dots, k$ . Also, the matrix, say  $S$ , of variances and covariances of the surveys estimates is available. As in Binder and Dick (1989, 1990), among several others, the  $\bar{y}_t$ 's may be modelled by

$$\bar{y}_t = \mu_t + e_t, \quad (16)$$

where  $e_t$  is the survey error at time  $t$  with variance-covariance matrix estimated by  $S$ . The means on each occasion,  $\mu_t$  for occasion  $t$ , follow an ARMA process. Model (16) is a special case of the random coefficients regression model, so that the appropriate marginal likelihood is different from (5).

A marginal or approximate conditional likelihood for correlation parameters in a random coefficients regression model is obtained as follows. Suppose in model (2) that  $\beta$  is a random vector modelled by  $\beta = W\delta + u$ , where  $W$  is a  $p \times q$  matrix of known values,  $\delta$  is a  $q \times 1$  vector of parameters, and  $u \sim N(0, \gamma^2\Gamma)$ , independent of  $\epsilon$ . Under the composite model  $y = XW\delta + Xu + \epsilon$ , the log-likelihood for  $\delta, \Omega, \Gamma, \gamma^2$ , and  $\kappa = \sigma^2/\gamma^2$ , denoted by  $L(\delta, \kappa, \gamma^2, \Gamma, \Omega)$ , is given by (3), with  $\Omega$  replaced by  $\kappa\Omega + X\Gamma X^T$  and  $X\beta$  replaced by  $XW\delta$ . Likewise, the marginal likelihood, denoted by  $L_M(\kappa, \Gamma, \Omega)$ , is given by (5), with  $X$  replaced by  $XW$  and  $\Omega$  replaced by  $\kappa\Omega + X\Gamma X^T$ . This yields

$$L_M(\kappa, \Gamma, \Omega) = \{ |\kappa\Omega + X\Gamma X^T|^{1/2} | (XW)^T(\kappa\Omega + X\Gamma X^T)^{-1}XW |^{1/2} g^{m-q} \}^{-1}, \quad (17)$$

where

$$g = y^T(\kappa\Omega + X\Gamma X^T)^{-1}y \\ - y^T(\kappa\Omega + X\Gamma X^T)^{-1}XW((XW)^T(\kappa\Omega + X\Gamma X^T)^{-1}XW)^{-1}(XW)^T(\kappa\Omega + X\Gamma X^T)^{-1}y.$$

Now the dimension of  $\Omega$  may be large in comparison to  $\Gamma$ ; this can be the case in sampling on successive occasions. As an alternate approach, one could take the likelihood implied by (3), multiply it by the distribution for  $\beta$ , and integrate over  $\beta$  to obtain the likelihood for the parameters under the random coefficient model. This will yield matrices of the same dimension as  $\Gamma$ .

Since  $S$  is available, an estimate of  $\Omega$ , the correlation matrix of the survey error, may be easily obtained. An estimate of  $\kappa = \sigma^2/\gamma^2$ , may also be obtained. From assumptions which lead to the marginal likelihood in (17), it is necessary to assume that  $e_t$  in (16) is a stationary random variable. Then an estimate of  $\sigma^2$  is the average of the diagonal elements in  $S$ . If  $\gamma^2$  is the variance of the  $\mu$ 's then the variation between  $\bar{y}_t$ ,  $t = 1, \dots, k$  provides an estimate of  $\sigma^2 + \gamma^2$ . From these two estimates, an estimate of  $\kappa$  may be obtained. Under model (16),  $X$  in (17) is the  $k \times k$  identity matrix, while  $W$  is a  $k \times 1$  column vector of 1's. The resulting marginal likelihood is a pseudo likelihood since some of the parameters have been replaced by estimates. In this case, the pseudo marginal likelihood for the parameters in  $\Gamma$  (pseudo since  $\kappa$  and  $\Omega$  have been replaced by their estimates) and is given by (17) with the appropriate substitutions. The parameters in  $\Gamma$  are the correlation parameters in the ARMA process on  $\mu_t$ . If  $k$ , the number of occasions, is relatively large in comparison to the number of parameters in  $\Gamma$ , then the marginal and approximate conditional likelihood estimates should be similar to the maximum likelihood estimator. For ease of computation, it seems that the full likelihood approach using the state space models as outlined by Binder and Dick (1989a, Section 3) appears to be the simplest approach to use in this situation.

## ACKNOWLEDGEMENTS

This work was supported by a grant from the Natural Sciences and Engineering Research Council of Canada. The author would like to thank the referee for his comments on an earlier draft of the paper.

## REFERENCES

- BELLHOUSE, D.R. (1978). Marginal Likelihoods for distributed lag models. *Statistische Hefte*, 19, 2-14.
- BELLHOUSE, D.R. (1989). Optimal estimation of linear functions of finite population means in rotation sampling. *Journal of Statistical Planning and Inference*, 21, 69-74.
- BELLHOUSE, D.R. (1990). On the equivalence of marginal and approximate conditional likelihoods for correlation parameters under a normal model. *Biometrika*, 77, 743-746.
- BINDER, D.A., and HIDIROGLOU, M.A. (1988). Sampling in time. *Handbook of Statistics, Volume 6 (Sampling)* (Eds. P.R. Krishnaiah and C.R. Rao). Amsterdam: North-Holland, 187-211.
- BINDER, D.A., and DICK, J.P. (1989). Modelling and estimation for repeated surveys. *Survey Methodology*, 15, 29-45.
- BINDER, D.A., and DICK, J.P. (1990). A method for the analysis of seasonal ARIMA models. *Survey Methodology*, 16, 239-253.
- BLIGHT, B.J.N., and SCOTT, A.J. (1973). A stochastic model for repeated surveys. *Journal of the Royal Statistical Society, Series B*, 35, 61-68.
- COCHRAN, W.G. (1977). *Sampling Techniques*. 3<sup>rd</sup> Ed. New York: Wiley.
- COX, D.R., and REID, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion). *Journal of the Royal Statistical Society, Series B*, 49, 1-39.
- CRUDDAS, A.M., REID, N., and COX, D.R. (1989). A time series illustration of approximate conditional likelihood. *Biometrika*, 76, 231-237.
- CUNIA, T., and CHEVROU, R.B. (1969). Sampling with partial replacement on three or more occasions. *Forest Science*, 15, 204-224.
- FRASER, D.A.S. (1967). Data transformations and the linear model. *The Annals of Mathematical Statistics*, 38, 1456-1465.
- HARVEY, A.C., and PHILLIPS, G.D.A. (1979). Maximum likelihood estimates of regression models with autoregressive-moving average disturbances. *Biometrika*, 66, 49-58.
- KALBFLEISCH, J.D., and SPROTT, D.A. (1970). Application of likelihood methods to models involving large numbers of parameters. *Journal of the Royal Statistical Society, Series B*, 32, 175-208.
- KILPATRICK, D.J. (1981). Optimum allocation in stratified sampling of forest populations on successive occasions. *Forest Science*, 27, 730-738.
- PATTERSON, H.D. (1950). Sampling on successive occasions with partial replacement of units. *Journal of the Royal Statistical Society, Series B*, 12, 241-255.
- ROBERTS, G., RAO, J.N.K., and KUMAR, S. (1987). Logistic regression analysis of sample survey data. *Biometrika*, 74, 1-12.
- TUNNICLIFFE WILSON, G. (1989). On the use of marginal likelihood in time series model estimation. *Journal of the Royal Statistical Society, Series B*, 51, 15-27.



# Sample Rotation and Estimation in the Survey of Employment, Payroll and Hours

IOANA SCHIOPU-KRATINA and K.P. SRINATH<sup>1</sup>

## ABSTRACT

The current Survey of Employment, Payroll and Hours, conducted by the Labour Division of Statistics Canada is a major monthly survey collecting data from a large sample of business establishments. This paper describes the methodology of the survey. The description of the stratification, sample size determination and allocation procedures is brief, whereas the description of the rotation procedure is more detailed because of its complexity. Some of the possible simplifications of the design are also highlighted.

KEY WORDS: Establishment; Response burden; Sampling frame.

## 1. INTRODUCTION

### 1.0 Objectives of the Survey

The Survey of Employment, Payroll and Hours (SEPH) is a monthly establishment based survey conducted by Statistics Canada.

The main objectives of SEPH are:

- (i) to provide monthly estimates of the total number of paid employees, average weekly earnings, average hourly earnings, average weekly hours and other related variables at the industry division-province level.
- (ii) to provide the above estimates for Canada at the three digit Standard Industrial Classification (SIC) level.
- (iii) to provide standard errors of all the estimates produced.

It is also intended to produce estimates at the three digit SIC-province level annually.

The survey covers all industries with the exception of agriculture, fishing and trapping, private household services, religious organizations and military services. For a detailed description of the objectives and uses of SEPH, see Cottrel-Boyd *et al.* (1980).

This article describes the sample selection and rotation as well as the estimation procedure adopted for the survey. Chapter 2 presents the sample selection and rotation procedure in detail. Chapter 3 is devoted to the estimation procedure. Some of the details relating to Chapter 2 are given in the Appendix. The Appendix also presents a simplified estimator of the number of live units.

For a complete description of the SEPH methodology, see Schiopu-Kratina and Srinath (1986).

### 1.1 Preliminary Definitions

Some of the terms used in this article are defined here for convenience.

- (i) Establishment – An establishment is the smallest unit that is a separate operating entity capable of reporting all elements of basic industrial statistics. The establishment is the statistical unit for SEPH. We will use the term unit for establishment.

<sup>1</sup> Ioana Schiopu-Kratina and K.P. Srinath, Business Survey Methods Division, Statistics Canada, Ottawa, K1A 0T6.

- (ii) Employment reporting unit (ERU) – For purposes of detailed geographical statistics, the establishment is often sub-divided into reporting units based mainly on location and sometimes on other considerations like payroll, *etc.*
- (iii) Standard Industrial Classification (SIC) – 1970. Each establishment is assigned a Standard Industrial Classification (SIC) code according to the nature of its activity. These SIC codes are defined in the SIC Manual (Statistics Canada 1970). For the purpose of the survey, 16 industry divisions (groups) which are groupings of specific three digits SIC's have been created.

In SEPH, the total number of paid employees associated with a unit is the characteristic chosen as a measure of size of that unit.

There are four size groups in SEPH and their boundaries are defined as follows: 0-19 for size group 1, 20-49 for size group 2, 50-199 for size group 3 and 200 or more for size group 4.

- (iv) A super-stratum is defined by an industry division, province and size group. With 16 industry divisions, 12 provinces and territories and 4 size groups, there are 768 super-strata.
- (v) A stratum is defined as a three digit SIC, province and size group. It is the finest level of detail for which estimates are obtained.
- (vi) The take-all portion of the population consists of units which are all included in the sample with certainty. It contains units in size group 4 and pre-specified units of the population. The take-some portion of the population consists of the remaining units which are subject to sampling as described in the following sections.

## 2. SAMPLE SELECTION AND ROTATION

### 2.0 Sample Size Determination and Allocation Procedure

In SEPH, the take-some sample size is determined at the industry group – province level based on a designed coefficient of variation of the estimate of the total number of employees for that industry group-province. The required sample size and the sampling fractions are calculated at the super-stratum level, using  $X$ -proportional allocation, where  $X$  represents the total number of employees. The sampling fractions are held constant from one month to the next. Details about the allocation procedure can be found in Schiopu-Kratina and Srinath (1986).

The actual selection is made at the stratum level. Due to the minimum sample requirements at that level, the number of sampled units is larger than the required sample size at the industry group – province level (see (2.2)).

### 2.1 Sample Selection

Let us now consider a specific stratum. Let  $N$  be the size of the take-some portion of the population and  $n$  the size of the take-some sample in this stratum.

Whereas the allocation of the sample to the super-strata is  $X$ -proportional, the allocation at the stratum level within each super-stratum is essentially proportional to the total number of units in the take-some portion of that stratum.

The sampling fraction in each stratum is given by the formula:

$$f = \max\left(f', \frac{1}{100}\right), \quad (2.1)$$

where  $f'$  is calculated at the corresponding super-stratum level. In order to reduce the instability of the estimates caused by small values of the sampling fraction, it was decided to set  $1/100$  as a minimum sampling fraction for all strata.

The detailed calculations of the sample size at stratum level are given in section 2.3 (see the derivation of the formula (2.8)). A systematic sample is drawn from each stratum.

## 2.2 The Rotation Scheme

The sample rotation (partial periodic replacement of the sample) in SEPH is designed primarily to reduce the response burden. From previous surveys, it appeared that the average response rate in strata in which there was no rotation was significantly lower than the average response rate in strata in which there was rotation. Also, the existence of a large portion of units common to the sample for two consecutive months improves the reliability of the estimates of month-to-month change relative to the estimate of change based on two independent monthly samples. Rotation of the sample in each stratum has to be done under certain constraints such as keeping the units out of the sample for a certain period of time after they rotate out of sample.

The monthly sample consists of 14 groups numbered from 0 to 13. Group 0 contains the take-all units of the stratum. Groups 1 to 13 are called rotation groups. The labels 1 to 12 on rotation groups indicate the month in which the units other than births rotated into the sample. For example, rotation group 1 contains mostly units which entered the sample in January and births, rotation group 2 contains mostly units which entered the sample in February and births, *etc.* Rotation group 13 contains units which have completed 12 months in the sample. These units are the oldest in terms of the time spent in the sample and are eligible to be rotated out. Each month, births are selected and allocated at random to the rotation groups.

At the time of the monthly selection and rotation, all units in the reference month are transferred to rotation group 13. In February, for example, all units in the rotation group 2 are transferred to rotation group 13. A replacement group is selected from "eligible for selection" units, and newly recorded units (births). The units of the replacement group (with the exception of 11/12 of all births) are then placed in rotation group 2, and they are not eligible to rotate out for at least 12 months. If sufficient units are available for a replacement group, the contents of group 13 are removed from the sample and are not eligible for reselection for 12 months. Otherwise, some units in group 13 are retained in the sample until such time that there are enough available units outside the sample to form a replacement group. This is done in order to maintain the minimum sample size or attain a sample size large enough to provide estimates with prespecified reliability. This way, in general at least 11/12 of the units stay in the take-some portion of the sample for two consecutive months.

The units that have left the sample are assigned to a waiting group which is divided into subgroups. A subgroup consists of units which were all removed from the sample in the same month. The waiting group contains 12 subgroups in every stratum. The time each unit has spent outside the sample is thus recorded to ensure that the units will not be reselected for at least 12 months. The units that have spent the required amount of time in the "not eligible for selection" group are transferred to the "eligible for selection" group and are thus assigned a positive probability of reselection.

To summarize, the entire take-some population at any given time consists of four groups of units. These are:

- (i) units that are in the sample for that month;
- (ii) units that are eligible for selection (E.F.S.);
- (iii) units in the waiting group which have rotated out of the sample less than 12 months ago and which are not eligible for selection (N.E.F.S.);
- (iv) births, *i.e.* units that have not been previously recorded on the frame.



The process of monthly selection and rotation involves an exchange of units among these groups. Some units leave group (i) for group (iii) and new units enter group (i) from group (ii), after some selected births from group (iv) have been transferred to group (ii). The remainder of the births are allocated to groups (ii) and (iii) after selection. This is done in order to insure that the sample is representative of the population in any given month.

## 2.3 Determination of the Sample Size and Weights

### 2.3.1 Monthly Updates

The sampling frame contains a large number of units which are inactive, out of business, out of scope *etc.* Apart from the burden of retaining an increasing number of inactive units on the frame, the estimators based on samples drawn from such a population are likely to have a high variance, due to the fact that the sample contains a high proportion of zero observations. Ideally, all such units should be eliminated from the sampling frame before the monthly sample is drawn. The frame is updated each month, after a monthly selection and rotation and prior to the next. For this reason, the indices we use to denote births and deaths on the frame are one unit higher than those used for the sample size in the sample selection preceeding the update. For example, after the initial sample selection, say  $n(0)$  units are in the sample, of which  $d(1)$  units are subsequently found to be dead units. Then  $D(1)$  denotes the number of dead units in the out-of-sample portion of the population and  $B(1)$  the number of units registered as births that month. In calculating the required sample size for the following month  $n(1)$ , one must take into account these updates (see (2.3)) as well as the size of the population at the time of the first sample selection  $N(0)$ .

### 2.3.2 Determination of Sample Size

The population of a given take-some stratum is a function of time and it will be denoted by  $N(t)$  say, where  $t$  is a positive integer which increases by one unit from one month to the next. The required sample size is, for each month:

$$n'(t) = [fN(t) + 0.5]. \quad (2.2)$$

Here  $[a]$  is the largest integer number which is not greater than  $a$ . The constant 0.5 is used for a better approximation in the rounding off procedure.

Suppose  $d(t)$  units are eliminated from the sample (in-sample deaths) and  $D(t)$  from the rest of the population of the stratum (out-of-sample deaths). Also let  $B(t)$  new units be recorded during the same time interval (births).

As a result, the size of the population of the cell at the time of the  $t^{\text{th}}$  selection is:

$$N(t) = N(t-1) - d(t) - D(t) + B(t). \quad (2.3)$$

Since the updates are not exhaustive, undetected inactive (dead) units are expected to exist in the population.

Let  $n_t(t)$  be the number of live units left in the previous month sample (after the updates) *i.e.*:

$$n_t(t) = n(t-1) - d(t). \quad (2.4)$$

We assume that there are no undetected dead units in the sample at this point. We can think of the population of a stratum as consisting of two domains: the domain of live units and the domain of dead units. The size of the dead domain is not known, but an estimate  $\hat{U}_d(t)$  can be calculated based on the information given by the sample and the updates (see Appendix). Let  $\hat{U}_d(t)$  be an estimate of the number of undetected dead units in the population at the time of the  $t^{\text{th}}$  monthly selection. Then:

$$N(t) = \hat{U}_l(t) + \hat{U}_d(t), \quad (2.5)$$

where  $\hat{U}_l(t)$  is the estimate of the number of live units.

The probability of choosing a dead unit when selecting a unit at random from the out-of-sample units is:

$$\hat{P}_d(t) = \min \left\{ \frac{\hat{U}_d(t)}{N(t) - n_l(t)}, 1 \right\}. \quad (2.6)$$

The required number of live units in the sample is:

$$n'_l(t) = f\hat{U}_l(t). \quad (2.7)$$

The replacement sample size is calculated in such a way as to ensure that the expected number of live units in the sample after selection is  $n'_l(t)$ .

Now assume that at the time of the  $t^{\text{th}}$  sample selection and rotation, of the  $n_l(t)$  live units in the sample,  $n_o(t)$  units are eligible to rotate out.

Since there are  $n_l(t) - n_o(t)$  live units left in the sample,  $n'_l(t) - n_l(t) + n_o(t)$  more live units are required in the sample for the  $t^{\text{th}}$  month.

In order to represent the births in the sample adequately,  $b(t) = fB(t)$  births should be selected at random and included in the sample.

Therefore:

$$\ell(t) = \max(n'_l(t) - n_l(t) + n_o(t) - b(t), 0),$$

live units should be selected from the eligible for selection group and added to the sample along with the selected births. Taking into account the existence of an unknown number of inactive units in the population and integerizing, it is required that:

$$n_i(t) = \min \left( \left\lceil \frac{\ell(t)}{1 - \hat{P}_d(t)} + 0.5 \right\rceil, n''(t) \right),$$

more units rotate into the same sample, with  $\hat{P}_d(t)$  given by (2.6) and  $n''(t) = N(t) - n_o(t)$ .

In calculating  $n_i(t)$ , we made the assumption that there are no inactive units among the births, so the expansion factor  $[1 - \hat{P}_d(t)]^{-1}$  is applied only to the "older" units in the E.F.S. group.

The sample size  $n(t)$  for the  $t^{\text{th}}$  month is:

$$n(t) = \max\{n_l(t) - n_o(t) + n_i(t) + b(t), m\}. \quad (2.8)$$

In (2.8),  $m$  represents the minimum required sample size for a stratum, which is presently set at 3. This additional requirement increases the sample size by 3,000 units in all strata, of which 1,800 are expected to be in the sample for a considerable length of time.

Of the  $n_i(t)$  units which rotate in,  $\hat{n}_d(t) = \hat{P}_d(t)n_i(t)$  are expected to be found inactive and  $\tilde{n}_i(t) = n_i(t) - \hat{n}_d(t)$  active. Thus, of the  $n(t)$  units in the sample after the  $t^{\text{th}}$  monthly selection and rotation,  $\hat{n}_i(t) = n_i(t) - n_o(t) + \tilde{n}_i(t) + b(t)$  are expected to be alive and they represent the  $\hat{U}_i(t)$  units of the live domain at the proper rate  $f$  (see (2.7) – (2.8)), when  $n(t) > m$  in (2.8).

### 2.3.3 Determination of Weights

The weight  $w(t)$  used for estimation for the  $t^{\text{th}}$  month is expressed in terms of the size of the population and sample at time  $t$ . However, the use of  $N(t)/n(t)$  as weight for estimation could lead to an overestimation of the live units in the population. Indeed,  $n(t)$  in formula (2.8) was chosen so that the expected number of live units in the sample equals the required sample size. The number of dead units in the sample drawn as described above may not represent the size of the dead domain at the proper rate. In (2.8),  $n_i$  is drawn from the general population and is thus expected to preserve the proportion between the dead and the live domain. No deaths are expected to be found among births and thus  $b(t)$  properly represents the birth subgroup of the population. There are, however,  $n_i(t) - n_o(t)$  units left in the sample from a previous selection, after rotation and the updates on the sample. The proportion of deaths among them is likely to be much smaller than the corresponding proportion in the general population, in spite of the fact that the updates are based on information from sources other than the survey. Then the value of  $N(t)/n(t)$  should be adjusted for the underrepresentation of dead units in the sample. This gives, when  $n(t) > m$ ,

$$\frac{N(t)}{n(t) + \hat{u}(t)} = \frac{1}{f}, \quad (2.9)$$

where  $\hat{u}(t)$  will be determined subsequently (see (2.10)). The value of  $\hat{u}(t)$  represents the “deaths” that have to be added to the sample to correctly represent the dead units in the population. Notice that when the first sample is drawn or if a redraw takes place, such an adjustment is not needed, *i.e.*  $\hat{u}(0) = 0$ .

In order to find a formula for  $\hat{u}(t)$ , we use (2.5) in the numerator of (2.9) and (2.8) in the denominator. By (2.7) – (2.8), we must also have:

$$\frac{\hat{U}'_d(t)}{\hat{n}_d(t) + \hat{u}(t)} = \frac{1}{f}.$$

With  $\hat{n}_d(t) = \hat{P}_d(t)n_i(t)$ , we obtain from above

$$\hat{U}'_d(t) = \frac{1}{f}[\hat{n}_d(t) + \hat{u}(t)] \quad \text{or} \quad \hat{u}(t) = f\hat{U}'_d(t) - \hat{n}_d(t). \quad (2.10)$$

The death adjustment is given by:

$$v(t) = \begin{cases} \hat{u}(t) & \text{if } \hat{u}(t) \geq -\hat{n}_d(t) \\ 0 & \text{if } \hat{u}(t) < -\hat{n}_d(t) \end{cases} \quad (2.11)$$



and the weight used in estimation is:

$$w(t) = \frac{N(t)}{n(t) + \hat{v}(t)}. \quad (2.12)$$

Note that the weight in (2.12) is defined using an estimate and so it is a random variable.

The use of the weight defined by (2.12) implies that the estimate of the number of live units in the population, defined by  $\hat{U}_t(t) = w(t)\hat{n}_t(t)$  does not exceed  $N(t)$ , the size of the population at the time of the  $t^{\text{th}}$  sample selection.

Let us define:

$$\hat{U}_d(t) = w(t) [\hat{v}(t) + \hat{n}_d(t)].$$

By (2.10) – (2.11), it follows that  $\hat{U}_d(t) \geq 0$  and its minimum value is 0 when  $\hat{v}(t) - \hat{n}_d(t) = 0$ . By (2.5), the maximum value of  $\hat{U}_t(t)$  is then  $N(t)$ .

The restriction that the estimator of live units be truncated at  $N(t)$  has implications on estimation which will be discussed in section 3.1. The estimator  $\hat{U}_d(t)$  is calculated recursively (see the Appendix) using 2.10 – 2.11 and the fact that  $\hat{v}(0) = \hat{u}(0) = 0$ .

It has to be noted that the formula (2.11) is slightly different from the formula giving the death adjustment in SEPH. Firstly, for the sake of simplicity, we did not consider here the cases when the minimum sample size  $m$  has to be used. In such cases, the use of the sampling fraction  $f$  in (2.10) is not appropriate. In SEPH, the previous month weight is used in (2.10) in lieu of  $f$  in all instances. Secondly, the death adjustment in SEPH is always taken to be positive. Formula (2.11) shows that it could be negative, as long as it is larger than  $-\hat{n}_d(t)$ . The actual instances in which this happens, or, more generally, when  $\hat{u} \leq 0$  are very rare.

The Appendix presents a formula for the estimator of the live units which does not require the use of the death adjustment.

## 2.4 Sampling of Births

As mentioned previously, every month new units are added to the frame. Since it is believed that these new units (births) may differ from the “old” units, a special birth strategy was designed, aimed at adequately representing the births in the sample.

Ideally, if  $B$  births are added during the current month and if  $f$  is the stratum sampling fraction, then  $b = fB$  births should be selected in the sample during that month. The selected births are randomly assigned to the rotation groups described in section 2.2. This ensures the same probability of rotating out of the sample for births as for the “old” units, so that the age distribution of the in-sample units is the same as the out-of-sample units.

Using the notation of the previous section, let  $n_i$  be the number of units required to rotate in at the time of the monthly sample selection excluding births and  $N'$  the number of units in the E.F.S. group (group (ii) of section 2.2). The birth strategy consists of a two-phase selection procedure. This procedure involves the formation of a common pool of births and “older” units from which a sample is then drawn. This was thought necessary because usually, the birth group is too small for sampling births separately each time. There are two ways of forming the common pool depending on the sizes of the birth group and the E.F.S. group. If:

$$\frac{n_i}{N'} > f, \quad (2.13)$$

then  $b'$  births are preselected from the birth group and a common pool of size  $(N' + b')$  is formed where:

$$b' = \frac{b N'}{n_i}. \quad (2.14)$$

Inequality (2.13) ensures that  $b' \leq B$  which means that the birth group is large enough and the preselection can take place. From the common pool of  $N' + b'$  units,  $n_i + b$  units are selected next and added to the sample.

The choice of  $b'$  as given by (2.14) ensures that the expected number of births in the sample is the desired one, since the probability of selecting one birth from the pool is  $b' / (b' + N')$  and therefore the expected number of births when  $n_i + b$  units are selected without replacement from this pool is  $(n_i + b) b' / (b' + N')$  which by (2.14) equals  $b$ . Similarly, it is easy to see that the expected number of "older" units is  $n_i$ .

In the complementary situation, when (2.13) does not hold, a common pool of size  $(n' + B)$  units is formed where:

$$n' = n_i / f \leq N'. \quad (2.15)$$

Then  $n'$  "older" units are selected from the E.F.S. group. Let us note that in this situation  $b'$  as given by (2.14) is larger than  $B$  and so the first procedure cannot be applied.

We now calculate the expected number of "old" units in the sample. Since the probability of selecting one "old" unit is now  $n' / (n' + B)$  and  $n_i + b$  units are drawn from the pool of  $n' + B$  units and placed in the sample, the expected number of "old" units is  $(n_i + b) n' / (n' + B)$  which equals  $n_i$ . The expected number of births is  $b$ .

It has to be noticed that an underrepresentation of births may occur in some situations. For example, if in some stratum no units are required to rotate in, then for the month in question the births will not be represented in the sample taken from that stratum. However this situation usually arises when the population size is small and in the long run, the representation of births can be expected to average out correctly.

The  $b$  births actually selected are randomly assigned to the rotation groups 1-12 in the sample, resulting, on the average, in  $b/12$  births assigned to each of these rotation groups. This ensures that the probability of a birth rotating out is the same as that of any other unit.

In order to keep the age distribution of the units in the groups (i) – (iii) (see section 2.1) constant, the non-selected births will be allocated to the E.F.S. and the N.E.F.S. group at random. That is, if  $N'$  is the number of units in the E.F.S. group and  $N''$  is the number of units in the N.E.F.S. group, then  $N' (B - b) / (N' + N'')$  unselected births will be assigned to the E.F.S. group and  $N'' (B - b) / (N' + N'')$  to the N.E.F.S. group.

### 3. ESTIMATION

#### 3.0 Introduction

In this chapter we describe the procedure for estimating the characteristic "the total number of paid employees".

As indicated in section 2.0, only the reliability of the estimates of the total employment at the industry group-province level of aggregation is prespecified. The estimates of characteristics other than the total employment have varying degrees of reliability. For example, estimates of average weekly earnings are expected to have higher reliability than the total employment.

The finest level of aggregation at which the estimates are published is the SIC-province level, but the basic "building blocks" for producing the estimates are the strata.

An outlier in SEPH is an observation in the take-some portion of the sample which is larger than a prespecified value.

The weight of each stratum is first calculated as in section 2.3, then it is adjusted for outliers in that stratum. Estimates of the total employment are computed for each stratum using the adjusted weights. There is no adjustment for nonresponse, as values for the nonrespondents are imputed. The estimate of the total employment for each stratum is obtained by adding the following totals:

- (i) the total employment for take-all units
- (ii) the total employment for outlier units
- (iii) the sum of the weighted values of employment for take-some units, excluding outliers.

Since the weight assigned to each outlier is one, outliers are treated as take-all units for the purpose of estimation and are therefore not used in the variance estimation.

### 3.1 Estimation of the Total of a Characteristic

Let us consider a specific stratum for a given month of the survey. Let  $N$  be the size of the take-some portion of the population of the stratum and  $n$  the number of take-some units in the sample for that month. If  $\hat{v}$  is the death adjustment (see (2.11)), then the original weight assigned to each unit in the sample for the purpose of estimation is (see (2.12)):

$$w = \frac{N}{n + \hat{v}}. \quad (3.1)$$

However, if  $t$  outliers are present in the sample with  $t \geq 1$  then this weight is modified by giving each outlier a weight of 1 and assigning to the remaining units in the take-some portion of the sample the weight:

$$w' = \frac{N - t}{n + \hat{v} - t}. \quad (3.2)$$

Let  $S$  represent the set of in-sample units. If  $Y(u)$  represents the value of employment corresponding to the unit  $u$  in the sample, then the estimate of the total employment in the stratum is:

$$\hat{Y} = \sum_{u \in S} w(u) Y(u). \quad (3.3)$$

Where  $w(u) = 1$  if  $u$  is an outlier or a take-all unit and  $w(u) = w'$  for all other units in the sample (see (3.2)).

Estimates of totals at any level of aggregation higher than the stratum are obtained adding the estimates of the stratum totals, for all strata in the level of aggregation considered.

Since for the purpose of estimation the outliers may be considered take-all units, we may replace, for the sake of simplicity  $w'$  by  $w$ , with  $w$  given by (3.1). Then  $N$  and  $n$  should be modified accordingly (see (3.2)).

Let  $N_\ell$  be the size of the live domain in the population of the stratum and  $n_\ell$  the number of live units in the sample. Let  $\bar{Y}_\ell$  represent the average employment of the live units in the sample.



Since only the live units contribute to the total employment, an estimate of the cell total is:

$$\hat{Y}_\ell = \hat{U}'_\ell \bar{Y}_\ell. \quad (3.4)$$

We consider the case  $fN > m$ .

In (3.4),  $\hat{U}'_\ell$  is an estimate of  $N_\ell$ , the number of live units in the population and is given by:

$$\hat{U}'_\ell = \frac{1}{f} n_\ell. \quad (3.5)$$

In (3.5),  $f$  is the stratum sampling fraction which is held fixed. The estimator based on the values of  $\hat{U}'_\ell$  is unbiased, that is:

$$E(\hat{U}'_\ell) = N_\ell. \quad (3.6)$$

As a consequence of the unbiasedness of the estimator based on (3.5),  $\hat{U}'_\ell$  may exceed  $N$  in some instances, as low possible outcomes of  $\hat{U}'_\ell$  compensate for it.

In SEPH, the estimate of live units  $\hat{U}_\ell$  is defined by:

$$\hat{U}_\ell = wn_\ell \quad (3.7)$$

with the weight given by (3.1).

The definition of the weight in SEPH (see the end of section 2.3) implies that:

$$\hat{U}_\ell = \begin{cases} \hat{U}'_\ell & \text{if } n_\ell \leq fN \\ N & \text{if } n_\ell > fN. \end{cases} \quad (3.8)$$

The estimate of the total employment in SEPH is defined by:

$$\hat{Y}_\ell = \hat{U}_\ell \bar{Y}_\ell. \quad (3.9)$$

An estimate of the total employment based on (3.5) is:

$$\hat{Y}'_\ell = \hat{U}'_\ell \bar{Y}_\ell. \quad (3.10)$$

The estimator based on (3.8) is biased and consequently the estimator of the total employment in SEPH is also biased.

However, the mean square error of the estimator based on (3.9) conditioned on  $n_\ell$  is smaller than the mean square error of the estimator based on (3.10), conditioned on  $n_\ell$ . We now sketch a proof of this claim.

It is not difficult to see that, for each particular outcome, the bias  $B_\ell$  of the estimator based on (3.9) and conditioned on  $n_\ell$  is given by:

$$B_\ell = (\hat{U}_\ell - N_\ell) \bar{Y}_\ell. \quad (3.11)$$

Similarly, we obtain for the estimator based on (3.10):

$$B'_\ell = (\hat{U}'_\ell - N_\ell) \bar{Y}_\ell. \quad (3.12)$$

We show that the conditional mean square error of the estimator (3.9) is smaller than the conditional mean square error of the estimator (3.10). The same result then holds for the unconditional mean square errors. We condition on the realized sample size of live units  $n_\ell$ . From (3.8) - (3.10),  $\text{Var}[\hat{U}'_\ell \bar{Y}_\ell | n_\ell] - \text{Var}[\hat{U}_\ell \bar{Y}_\ell | n_\ell] = [n_\ell^2 f^{-2} - N^2] 1\{n_\ell > fN\} \text{Var}[\bar{Y}_\ell | n_\ell]$ . Notice that  $n_\ell^2 f^{-2} - N^2 > 0$  on the set  $\{n_\ell > fN\}$ . We now compare  $[B'_\ell]^2$  and  $B_\ell^2$ :

$$[B'_\ell]^2 - B_\ell^2 = 1\{n_\ell f^{-1} > N\} \{[\hat{U}'_\ell - N_\ell]^2 \bar{Y}_\ell^2 - (N - N_\ell)^2 \bar{Y}_\ell^2\}.$$

But  $\hat{U}'_\ell - N_\ell = f^{-1}n_\ell - N_\ell > N - N_\ell$  if  $n_\ell f^{-1} > N$ .

Therefore  $[B'_\ell]^2 - B_\ell^2 \geq 0$ . Since  $\text{MSE}[\hat{U}'_\ell \bar{Y}_\ell | n_\ell] = \text{Var}[\hat{U}'_\ell \bar{Y}_\ell | n_\ell] + [B'_\ell]^2$  and  $\text{MSE}[\hat{U}_\ell \bar{Y}_\ell | n_\ell] = \text{Var}[\hat{U}_\ell \bar{Y}_\ell | n_\ell] + [B_\ell]^2$ , the term-by-term comparison leads to the conclusion that:

$$\text{MSE}[\hat{U}'_\ell \bar{Y}_\ell | n_\ell] \geq \text{MSE}[\hat{U}_\ell \bar{Y}_\ell | n_\ell].$$

This important property motivates the choice of the estimator (3.9) over (3.10) for the total employment in SEPH.

## APPENDIX

In this Appendix, we use the notation of section 2.3.2.

We first derive the formula for  $\hat{U}_d(t)$ , the size of the dead domain used in SEPH for the  $t^{\text{th}}$  selection.

Recall  $\hat{U}_d(t) = w(t)[\hat{v}(t) + \hat{n}_d(t)]$ . At the time of the  $t^{\text{th}}$  update,  $d(t+1)$  dead units are found in the sample. We can replace therefore  $\hat{n}_d(t)$ , the estimated number of dead units in the sample by  $d(t+1)$  in order to obtain an update of  $\hat{U}_d(t)$ , namely  $\hat{N}_d(t+1)$ :

$$\hat{N}_d(t+1) = w(t)[\hat{v}(t) + d(t+1)]. \quad (1.1)$$

Formula (1.1) uses the death adjustment from the previous month. The initial value of  $\hat{v}$  is  $\hat{v}(0) = 0$  (see the remark after (2.9) and the definition of  $\hat{u}$ ) and  $w(0) = f^{-1}$ . Now the estimate of the size of the dead domain for the  $(t+1)^{\text{th}}$  monthly selection is:

$$\hat{U}_d(t+1) = \max(\hat{N}_d(t+1) - D(t+1) - d(t+1), 0). \quad (1.2)$$

Notice that  $\hat{U}_\ell(t+1)$  can be calculated from (2.5) when  $\hat{U}_d(t+1)$  is known and vice versa. An alternative form for  $\hat{U}_\ell(t+1)$  is obtained recursively as follows. Let us assume that  $\hat{U}_\ell(t)$  is known before the  $t+1^{\text{th}}$  selection of the sample (recall that  $t=0$  is used for the first, or original selection). Then  $\hat{U}_d(t)$  is also known and can be used to calculate  $\hat{P}_d(t)$ , the probability of selecting a dead unit from the out-of-sample units (see formula 2.6). This probability is then used to calculate the required number of units which should rotate in as described in 2.3.2, as well as the expected number of live units in the sample at the time of the  $(t+1)^{\text{th}}$  selection,  $\tilde{n}_\ell(t)$ .

Then the weight used in estimation for the next selection is  $\hat{U}_\ell(t)/\tilde{n}_\ell(t)$ . After selection, the  $(t+1)^{\text{th}}$  update takes place and the actual number of live units in the sample is found to be  $n_\ell(t+1)$ . The estimate of the size of the live domain for the following selection can be calculated

$$\hat{U}_\ell(t+1) = \min \left\{ \frac{\hat{U}_\ell(t)}{\tilde{n}_\ell(t)} n_\ell(t+1) + B(t+1), N(t+1) \right\}$$

and so forth. To initiate the process, note that the weight used in the first estimation is  $w(0) = f^{-1}$  and after the first update,

$$\hat{U}_\ell(1) = \min \{ w(0) \times n_\ell(1) + B(1), N(1) \}.$$

## REFERENCES

- COTTREL-BOYD, T.M., DUNN, M.R., HUNTER, G.E., and SRINATH, K.P. (1980). Development of the redesign of the Canadian establishment based employment surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 8-15.
- SCHIOPU-KRATINA, I., and SRINATH, K.P. (1986). The methodology of the Survey of Employment, Payroll and Hours. Working Paper No. BSMD-86-010E, Statistics Canada.
- STATISTICS CANADA (1970). *Standard Industrial Classification Manual*. Catalogue 12-501, Ottawa: Statistics Canada.



## Estimating a System of Linear Equations with Survey Data

PHILLIP S. KOTT<sup>1</sup>

### ABSTRACT

This paper develops a framework for estimating a system of linear equations with survey data. Pure design-based sample survey theory makes little sense in this context, but some of the techniques developed under this theory can be incorporated into robust model-based estimation strategies. Variance estimators with the form of the single equation “linearization” estimator are nearly unbiased under many complex error structures. Moreover, the inclusion of sampling weights in regression estimation can protect against the possibility of missing regressors. In some situations, however, the existence of missing regressors can make the estimation of a system of equations ambiguous.

**KEY WORDS:** Sampling weights; Putative missing regressor; Robust; Nearly unbiased.

### 1. INTRODUCTION

Kott (1991) showed that design-based techniques developed for estimating a single linear regression equation could be exploited in a more conventional model-based framework. In particular the use of sample weighted regression was shown to help protect against the possible existence of missing regressors, while the so-called linearization variance estimator was shown to produce nearly unbiased estimators of mean squared error for many complex variance structures.

This paper extends those results to the estimation of a system or “grouping” of linear equations, a topic of considerable interest to econometricians (see, for example, Johnston 1972, pp. 238-241). Two simple examples may shed some light onto the subject for those not already schooled in econometric methods or their equivalent.

Suppose we have a sample of farmers and want to estimate the relationship between the amount of planted soybean acres and the size of the farm. Zellner (1962) showed in effect that even if a simple quadratic equation with independent and identically distributed errors correctly described the universe, a better estimator than the one produced by ordinary least squares (OLS) might exist. This estimator could be found by taking into consideration other linear relationships, say between planted *corn* acres and farm size, that had errors terms correlated with those in the original relationship. Zellner called the system-wide estimation of a group of such equations “seemingly unrelated regression.” Oddly, in order for Zellner’s generalized least squares (GLS) estimator to produce different results from OLS, it is necessary for some equations to contain regressors not found in other equations. Alternatively, one can think of each equation as containing the same regressors but with certain coefficients constrained to zero.

A second example concerns a sample of firms each producing one output,  $y$ , from two inputs,  $x_1$  and  $x_2$ , with unit prices,  $p_1$  and  $p_2$ . Economists often assume that each firm possesses the same technology (plus or minus an error term). Given  $p_1$ ,  $p_2$ , and  $y$ , each firm would choose  $x_1$  and  $x_2$  so as to minimize total cost,  $c = p_1x_1 + p_2x_2$ . Suppose that the

<sup>1</sup> Phillip S. Kott, Special Assistant for Economic Survey Methods, U.S. Bureau of the Census, Room 3061-3, Washington, DC, 20233, USA.

relation between  $p_1$ ,  $p_2$ ,  $y$  and the cost minimizing  $c$  can be expressed by the following equation (on average):

$$\log(c) = b_0 + b_1 \log(p_1) + b_2 \log(p_2) + b_3 \log(y). \quad (1)$$

Economic theory tells us that a rational firm faced with implicit cost equation (1) would choose its level of  $x_1$  so that

$$x_1 p_1 / c = b_1. \quad (2)$$

Naturally, in order to estimate equations (1) and (2), we need to add a stochastic structure. For simplicity, assume that both equations (1) and (2) fit the behavior of all firms subject to respective independent (across firms) and identically distributed random errors. Observe that in addition to the strong possibility that the error terms in the two equation will be correlated for a particular firm, there is also a coefficient ( $b_1$ ) shared by both equations.

When faced with a system of linear equations in which the coefficients are known to be constrained, the design-based approaches to linear regression reviewed in Kott (1990a) make little sense. For that reason, although design-based *practice* inspires many of the procedures discussed here, only the extended model-based approach introduced in Kott (1991) will be used to justify them.

Section 2 lays out the theoretical model for the estimation of a system of linear equations based on data from the full population. Section 3 introduces the sample weighted analogues of full population OLS and GLS estimators for a system of linear equations. Section 4 addresses robust mean squared error estimation of both the sample weighted OLS and GLS estimators employing a straightforward generalization of the linearization variance estimator (see, for example, Shah, Holt, and Folsom 1977). Section 5 discusses a general method for developing test statistics that can be used to evaluate, among other things, whether sample weighted OLS and GLS are actually estimating the same thing. Section 6 explores a simple example. Section 7 sketches an extension of the methodology developed here to what econometricians call “simultaneous equations.” In the stochastic version of equation (1), for example, many economists believe that  $\log(y)$  should be treated as a random variable and that  $\log(c)$  can be assumed to be fixed. This causes a simultaneity bias if not specifically addressed by techniques like two and three stage least squares (see Johnston 1972, pp. 341-420). Finally, section 8 contains a brief discussion.

## 2. FULL POPULATION ESTIMATION

### 2.1 The Unconstrained System:

Suppose we have a population containing  $M$  data points. Each data point  $i$  is associated with  $G + \tilde{K}$  observed variables satisfying the following model:

$$Y = \tilde{X} \tilde{\beta} + U + V, \quad (3)$$

where  $Y$  is an  $M \times G$  matrix of observed dependent variables (the  $i$ th row of  $Y$  contains the dependent variables associated with the  $i$ th data point),

$\tilde{X}$  is an  $M \times \tilde{K}$  matrix of observed independent or regressor variables (the  $i$ th row of  $\tilde{X}$  contains the independent variables associated with the  $i$ th data point),

$\tilde{\beta}$  is an  $\tilde{K} \times G$  matrix of parameters,

$U$  is an  $M \times G$  matrix satisfying the relationship  $\lim_{M \rightarrow \infty} \tilde{X}' U / M = 0_{\tilde{K} \times G}$  (a matrix of zeroes) – this assumes that there is an underlying process generating the data points which could in principle generate points *ad infinitum* (see Kott 1991), and

$V$  is a  $M \times G$  matrix of random variables such that  $E(V) = 0_{M \times G}$  and  $E(v_{is} v_{it}) = \sigma_{st(i)}$ .

It is well known that if  $U \equiv 0_{M \times G}$ ,  $E(v_{is} v_{jt}) = 0$  for  $i \neq j$ , and  $\sigma_{st(i)} = \sigma_{st}$  for all  $i$ , then

$$\tilde{B}_{OLS} = (\tilde{X}' \tilde{X})^{-1} \tilde{X}' Y \quad (4)$$

is the best linear unbiased estimator for  $\tilde{\beta}$  (see, for example, Johnston 1972, p. 240). This means that the  $g$ th column of  $\tilde{B}_{OLS}$ , call it  $B_{.g}$ , is the best linear unbiased estimator of  $\beta_{.g}$ , where

$$y_{.g} = \tilde{X} \beta_{.g} + u_{.g} + v_{.g}, \quad (5)$$

and  $y_{.g}$ ,  $u_{.g}$ , and  $v_{.g}$  are the  $g$ th columns of  $Y$ ,  $U$ , and  $V$ , respectively. Equation (5) can be viewed as the  $g$ th equation in the system of equations represented by equation (3).

Let us call the matrix  $U$  in equation (3) the *putative missing regressor* matrix. Usually, in conventional (*i.e.*, model-based) regression analysis that part of the dependent variable (or variables) not capturable by a linear combination of the independent variables is (are) assumed to be purely random. Here, however, we follow Kott (1991) and allow for the possibility of non-random missing regressors. Note that even when  $U \neq 0_{M \times G}$ ,  $\tilde{B}_{OLS}$  remains nearly (*i.e.*, asymptotically) unbiased.

## 2.2 The (Possibly) Constrained System:

Efficient estimation is a more complicated matter when there are constraints on some elements of  $\tilde{\beta}$ ; for example, when  $\tilde{\beta}_{kg}$  is known to be zero or when  $\tilde{\beta}_{hg}$  is known to equal  $\tilde{\beta}_{hj}$ .

In this paper, we are interested in a (possibly) constrained systems of equations that can be modelled directly with the following equation:

$$y = X\beta + u + v, \quad (6)$$

where  $y = (y_{.1}', y_{.2}', \dots, y_{.G}')'$ ,  $u$  and  $v$  are defined in an analogous manner,  $X$  is an  $MG \times K$  matrix,  $\beta$  is a  $K \times 1$  vector, and  $K \leq G\tilde{K}$ . By definition,  $\lim_{M \rightarrow \infty} X' u / M = 0_K$ .

When the original  $\tilde{\beta}$  in equation (3) is unconstrained,  $K = G\tilde{K}$ , and

$$X = \begin{bmatrix} \tilde{X} & & \\ & \tilde{X} & \\ & & \ddots \\ & & & \tilde{X} \end{bmatrix}.$$

When the original  $\tilde{\beta}$  is constrained, however,  $K < G\tilde{K}$ . For example, when an element of  $\tilde{\beta}$  is known to be zero, it can be removed from the  $\beta$  vector in equation (6) along with the column of the  $X$  matrix that corresponds to it. When two elements in the same row of  $\tilde{\beta}$  are known to be equal, the second can be removed from  $\beta$ , and  $X$  can be adjusted accordingly (it will no longer be block diagonal).



When  $u \equiv 0_{MG}$  and  $\text{Var}(v) = \sum \otimes I_M$  (where  $\sum = \{\sigma_{st}\}$ ), then  $b_{OLS} = (X'X)^{-1}X'y$  is an unbiased estimator for  $\beta$ , but  $b_{GLS} = (X'[\sum^{-1} \otimes I_M]X)^{-1}X'[\sum^{-1} \otimes I_M]y$ , where  $I_M$  is the  $M \times M$  identity matrix, is the best linear unbiased estimator. In practice, the elements of  $\sum$  have to be estimated from the sample, say by  $\hat{\sigma}_{gf} = r_{\cdot g}'r_{\cdot f}/M$ , where  $r_{\cdot g} = y_{\cdot g} - X_{\cdot g}b_{OLS}$ .

It is well known that  $b_{OLS}$  and  $b_{GLS}$  are equal when the parameter matrix in (3) is unconstrained (again, see Johnston 1972, p. 240). Turning to the constrained case, if  $u \neq 0_{MG}$ , then both  $b_{OLS}$  and  $b_{GLS}$  are nearly unbiased estimators of  $\beta$  when  $\lim_{M \rightarrow \infty} \tilde{X}'U/M = 0_{K \times G}$  holds as we originally assumed. Unfortunately,  $b_{GLS}$  may not be nearly unbiased under the weaker assumption that  $\lim_{M \rightarrow \infty} X'u/M = 0_K$ , which is more in line with the extended model in Kott (1991) when (6) is viewed as a single equation.

To see why this is, let  $X_{\cdot g}$  denote the  $M \times K$  matrix formed from the  $\{(g-1)M+1\}$ th through the  $\{gM\}$ th row of  $X$  and  $\sum^{-1} = \{\sigma^{fg}\}$ , then

$$E(b_{GLS} - \beta) \propto X'[\sum^{-1} \otimes I_M]u/M =$$

$$\sum_g X_{\cdot g}' \left( \sum_f \sigma^{fg} u_{\cdot f} \right) / M = \sum_g \sum_f \sigma^{fg} X_{\cdot g}' u_{\cdot f} / M,$$

which approaches zero as  $M$  grows large under the stronger assumption but not necessarily the weaker one.

### 3. ESTIMATION WITH SURVEY DATA

Suppose now that we observe variables values for only a random sample of the population. Let  $P = \text{diag}\{p_i\}$ , where  $p_i$  is the probability of selection for data point  $i$ . Let  $S = \text{diag}\{s_i\}$ , where  $s_i = 1$  if data point  $i$  is in the sample and 0 otherwise. Finally, let  $W = (m/M)P^{-1}S = \text{diag}\{w_i\}$  be the matrix of sampling weights, where  $m$  is the sample size. When all the  $p_i = m/M$ , note that  $W = S$ .

It is not difficult to show that for many sample designs and populations (see Kott 1990b and 1991), the *sample weighted OLS estimator*:

$$\hat{\beta}_{W \cdot OLS} = (X'[I_G \otimes W]X)^{-1}X'[I_G \otimes W]y \quad (7)$$

is a design consistent estimator for  $b_{OLS}$ , which means that  $\text{plim}_{m \rightarrow \infty} (\hat{\beta}_{W \cdot OLS} - b_{OLS}) = 0_K$ . Under similar conditions, *sample weighted GLS estimator*:

$$\begin{aligned} \hat{\beta}_{W \cdot GLS} &= (X'[I_G \otimes W][\hat{\sum}^{-1} \otimes I_M]X)^{-1}X'[I_G \otimes W][\hat{\sum}^{-1} \otimes I_M]y \\ &= (X'[\hat{\sum}^{-1} \otimes W]X)^{-1}X'[\hat{\sum}^{-1} \otimes W]y, \end{aligned} \quad (8)$$

where

$$\hat{\sigma}_{gf} = r_{\cdot g}' W r_{\cdot f} / \sum_{i=1}^M w_i, \quad \text{and} \quad r = y - X\hat{\beta}_{W \cdot OLS},$$

is a design consistent estimator for  $b_{GLS}$ . Like  $b_{OLS}$  and  $b_{GLS}$  (and for the same reasons),  $\hat{\beta}_{W \cdot OLS}$  and  $\hat{\beta}_{W \cdot GLS}$  are equal for an unconstrained system of equations.

If  $\hat{\beta}_{W \cdot OLS}$  and  $\hat{\beta}_{W \cdot GLS}$  are design consistent, both are also nearly unbiased estimators of  $\beta$  when  $\lim_{M \rightarrow \infty} \tilde{X}'U/M = 0_{\tilde{K} \times G}$ , because  $b_{OLS}$  and  $b_{GLS}$  are; however,  $\hat{\beta}_{W \cdot GLS}$  - like  $b_{GLS}$  - may not be nearly unbiased under the weaker assumption that  $\lim_{M \rightarrow \infty} X'u/M = 0_K$ . (Unbiasedness here is always defined with respect to the model in equation (6)).

#### 4. MEAN SQUARED ERROR ESTIMATION

Suppose the sample design is such that there are  $H$  strata,  $n_h$  distinctly sampled PSU's in stratum  $h$ , and  $m_{hj}$  sampled data points in PSU  $hj$ . Both  $\hat{\beta}_{W \cdot OLS}$  and  $\hat{\beta}_{W \cdot GLS}$  have the form  $\hat{\beta} = Cy$ . Without loss of generality, they can be rewritten as  $\hat{\beta} = C^*y^*$ , where  $y^* = (y_{11}', \dots, y_{Hn_H}')$  contains only elements corresponding to sampled data points, and  $y_{hj}$  is the vector of  $G \times m_{hj}$   $y$ -values associated with data points in PSU  $hj$ . Define  $r^*$  and  $r_{hj}$  in an analogous manner to  $y^*$  and  $y_{hj}$ .

Let  $D_{hj}$  be a diagonal matrix of 0's and 1's such that  $D_{hj}y^* = (0', \dots, y_{hj}', \dots, 0')$ , and let  $g_{hj} = C^*D_{hj}r^*$ . Extending the design-based linearization variance estimator in a straight forward manner, the estimator for the mean squared error of  $\hat{\beta} = C^*y^*$  has the form:

$$mse = \sum_{h=1}^H \frac{n_h}{n_h - 1} \left[ \sum_{j=1}^{n_h} g_{hj} g_{hj}' - \frac{1}{n_h} \left( \sum_{j=1}^{n_h} g_{hj} \right) \left( \sum_{j=1}^{n_h} g_{hj}' \right) \right]. \tag{9}$$

Under mild restrictions on the sampling design, mse is nearly unbiased when  $U$  (from (3))  $\equiv 0_{M \times G}$  and  $V$  obeys the following property:

$$| E(v_{sg} v_{tf}) | \begin{cases} = 0 & \text{when } s \text{ and } t \text{ are from different PSU's} \\ < Q & \text{otherwise.} \end{cases}$$

See Kott (1991) for the proof in the  $G = 1$  case; the extension to the  $G > 1$  case is trivial. The estimator mse remains reasonable when  $U \neq 0_{M \times G}$  (see Kott 1990a).

#### 5. TEST STATISTICS

Let  $\hat{\beta}_{I \cdot OLS}$  and  $\hat{\beta}_{I \cdot GLS}$  be the unweighted counterparts of  $\hat{\beta}_{W \cdot OLS}$  and  $\hat{\beta}_{W \cdot GLS}$  derived by replacing the  $W$  in (7) and (8) by  $S$ . One is often interested in determining whether using the sampling weights really matters. This comes down to testing whether  $\hat{\beta}_{I \cdot OLS}$  and  $\hat{\beta}_{W \cdot OLS}$  are significantly different; that is, whether they are estimating the same thing.

When weights *are* determined to matter, another question of some interest is whether  $\hat{\beta}_{W \cdot OLS}$  and  $\hat{\beta}_{W \cdot GLS}$  are significantly different; that is, does  $\lim_{M \rightarrow \infty} \tilde{X}'U/M = 0_{\tilde{K} \times G}$  hold so that these two estimators are estimating the same thing?

A general statistic for testing whether:

$$\hat{\beta}_{(1)} = \sum_h \sum_j \{C_{(1)}^* D_{hj} y^*\} \quad \text{and} \quad \hat{\beta}_{(2)} = \sum_h \sum_j \{C_{(2)}^* D_{hj} y^*\}$$

are equal is

$$T^2 = [\hat{\beta}_{(1)} - \hat{\beta}_{(2)}]' A^{-1} [\hat{\beta}_{(1)} - \hat{\beta}_{(2)}], \tag{10}$$

where

$$A = \sum_{h=1}^H \frac{n_h}{n_h - 1} \left[ \sum_{j=1}^{n_h} d_{hj} d_{hj}' - \frac{1}{n_h} \left( \sum_{j=1}^{n_h} d_{hj} \right) \left( \sum_{j=1}^{n_h} d_{hj}' \right) \right],$$

$$d_{hj} = C_{(1)}^* D_{hj} r_{hj(1)} - C_{(2)}^* D_{hj} r_{hj(2)}, \text{ and } r_{hj(f)} = y_{hj} - X_{hj} \hat{\beta}_{(f)}.$$

Under the null hypothesis, the test statistic,  $T^2$ , is asymptotically a  $\chi^2$  random variate with  $K$  degrees of freedom. Given our concern for robustness, it seems prudent to question the null hypothesis when  $\text{prob}(\chi_{(K)}^2 > T^2)$  is at considerably less than the standard 0.1 or 0.05 level, but not when  $T^2$  is less than its expected value,  $K$ .

## 6. AN EXAMPLE

Consider the following example synthesized from data from the National Agricultural Statistics Service's June 1989 Agricultural Survey. The data set, previously analyzed in Kott (1990a), is briefly described below.

A sample of 17 primary sampling units was selected from among 4 strata. These PSU's were then subsampled yielding a total sample of 252 farms. Although the sample was random, not all farms had the same probability of selection.

Suppose we are interested in estimating the parameters,  $\beta_1$  and  $\beta_2$ , of the following equation:

$$y_{1i} = x_{1i}\beta_1 + x_{2i}\beta_2 + u_{1i} + v_{1i}, \quad (11)$$

where  $i$  denotes a farm,

$y_{1i}$  is farm  $i$ 's planted soybeans to cropland ratio when  $i$ 's cropland is positive, zero otherwise;

$x_{1i}$  is 1 if farm  $i$  has positive cropland, zero otherwise; and

$x_{2i}$  is farm  $i$ 's cropland divided by 10,000.

(Note: dropping all sampled farms with zero cropland from the regression equation will have no effect on the parameter estimation, but it can affect mean squared error estimation.)

Letting  $\hat{\beta}_{(1)}$  in equation (10) be the pure OLS estimator for the vector  $(\beta_1, \beta_2)'$ , and  $\hat{\beta}_{(2)}$  be the sample weighted estimator, one computes a  $T^2$  of 4.58. Under the null hypothesis that OLS and sample weighted least squares are estimating the same thing (for which  $u_{1i} \equiv 0$  is sufficient but not necessary),  $T^2$  is asymptotically  $\chi_{(2)}^2$ . We cannot reject this null hypothesis at the 0.1 level. Nevertheless, since  $T^2$  is considerably greater than 2, it seems that the existence of a putative missing regressor is more than likely. Thus, the sample weighted regression estimator should be employed rather than the OLS estimator.

Table 1 displays both the pure OLS and the sample weighted coefficient estimates. Although the sample weighted estimator for  $\beta_2$  is not significantly different from zero at the 0.1 level, we retain it in the model because it exceeds its estimated root mean squared error. This parallels the reasoning for preferring sample weighted regression over OLS.

Notice the loss of efficiency that results from using the sample weighted estimator in place of pure OLS. The estimated root mean squared error for the  $\beta_2$  estimator more than doubles (note: both root mean squared errors were estimated using equation (9)).



Table 1  
Alternative Estimates for Equation (11)

OLS	$y_{1i} = 0.268x_{1i} - 0.92x_{2i} + u_{1i} + v_{1i}$ (.044) (3.95)
sample weighted	$y_{1i} = 0.191x_{1i} + 12.15x_{2i} + u_{1i} + v_{1i}$ (.075) (9.95)
sample weighted GLS	$y_{1i} = 0.197x_{1i} + 10.26x_{2i} + u_{1i} + v_{1i}$ (0.71) (6.97)

Numbers in parentheses are root mean squared errors.

We can increase the efficiency of the sample weighted estimator by adding a second farm equation and estimating it and (11) as a system. Let

$$y_{2i} = x_{1i}\beta_3 + u_{2i} + v_{2i}, \tag{12}$$

where  $y_{2i}$  is farm  $i$ 's planted corn to cropland ratio when  $i$ 's cropland is positive, zero otherwise.

The sample weighted estimators in Table 1 and their estimated root mean squared errors are unchanged under system-wide sample weighted OLS. The system approach, however, allows us to calculate sample weighted GLS estimator for  $\beta_1$  and  $\beta_2$  which are also displayed in the table. Observe that the estimated root mean squared error for  $\beta_2$  is reduced by approximately 30% without a loss of robustness, assuming that sample weighted OLS and GLS are estimating the same thing.

The  $T^2$  value for a test comparing the sample weighted OLS and GLS estimators for the vector  $(\beta_1, \beta_2)'$  is 0.97. This number is considerably less than 2. Thus, the two estimators do appear to be estimating the same thing. That is to say, there is no additional regressor in one equation related to the putative missing regressor in the other (which is not surprising since when an  $x_{2i}$  term was added to the right hand side of equation (12), its estimated coefficient was less than its estimated root mean squared error).

7. SIMULTANEOUS EQUATIONS

In a simultaneous equation framework, some of the columns of the dependent variable matrix,  $Y$ , (see (3)) are actually contained on the right hand side of the  $g$ th equation (see (5)). Formally, we can write

$$y = Y_{(\cdot)}\alpha + X\beta + u + v \quad \text{or} \quad y = Z\delta + u + v,$$

where

$$Y_{(\cdot)} = \begin{bmatrix} Y_{(1)} \\ Y_{(2)} \\ \vdots \\ Y_{(G)} \end{bmatrix},$$

$$Z = (Y_{(\cdot)}X), \quad \text{and} \quad \delta = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}.$$

Most of the columns of  $Y_{(g)}$  are 0-vectors. The rest (no more than  $G-1$  columns) are columns of  $Y$  from equation (3).

Define  $\hat{Y}_{(g)}$  as  $\tilde{X}(\tilde{X}'W\tilde{X})^{-1}\tilde{X}'WY_{(g)}$ . Now replace  $X$  in (5) by  $\hat{Z} = (\hat{Y}_{(\cdot)} X)$  and proceed as before. Equation (7) produces  $\hat{\delta}_{W\cdot OLS}$ , akin to two stage least squares, while (8) produces  $\hat{\delta}_{W\cdot GLS}$ , akin to three stage least squares. Mean squared error estimation follows along the same line of reasoning that produced equation (9).

## 8. DISCUSSION

The purpose of this paper was to show how procedures developed in the design-based survey sampling literature – in particular, sample weighted regression and the linearization mean squared error estimator – could be adopted to the estimation of a system of linear equations.

One somewhat unexpected discovery was when estimating the parameters of a constrained linear system, the sample weighted analogues of OLS and GLS might be estimating different things. On further reflection this is not so suprising. If there are missing regressors in our working model, perhaps we don't always know enough about the true model to put constraints on the parameters in the first place.

It is important to realize that mse in equation (9) can be used to estimate the mean squared error of parameter estimators even when there are no missing regressors. The advantages of mse to conventional practice is that it allows for the possibility of heteroscedasticity and complex correlations across data points (but within PSU's).

If there are no missing regressors, however, the following estimator has all the advantages of mse and is generally more efficient:

$$mse' = \sum_{h=1}^H \sum_{j=1}^n \frac{n}{n-1} \{g_{hj}g_{hj}' - gg'/n\}, \quad (13)$$

where  $n = \sum n_h$  and  $g = \sum \sum g_{hj}$  (note: if  $Xq = (1, \dots, 1)'$  for some  $K$ -vector  $q$ , then  $g = 0$ ).

When there *are* missing regressors the diagonal elements of  $mse'$  may tend to be biased upward. The reasoning here follows that in Wolter (1985) for collapsed strata variance estimators in design-based sampling theory.

## REFERENCES

- JOHNSTON, J. (1972). *Econometric Methods*, (second edition). New York: McGraw Hill.
- KOTT, P.S. (1991). A model-based look at linear regression with survey data. *American Statistician*, forthcoming.
- KOTT, P.S. (1990a). What does performing a linear regression on survey data mean? *Proceedings of the Section on Survey Research Methods, American Statistical Association*, forthcoming.
- KOTT, P.S. (1990b). The design consistent regression estimator and its conditional variance. *Journal of Statistical Planning and Inference*, 24, 287-296.
- SHAH, B.V., HOLT, M.M., and FOLSOM, R.E. (1977). Inference about regression models from sample survey data. *Bulletin of the International Statistical Institute*, 47, 43-57.
- WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.
- ZELLNER, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association*, 57, 348-368.

## **The Evaluation of Errors in National Accounts Data: Provisional and Revised Estimates**

**LUIGI BIGGERI and UGO TRIVELLATO<sup>1</sup>**

### **ABSTRACT**

This article is a critical review of recent developments in the evaluation of the reliability of provisional national accounts estimates. First, we will sketch a theoretical outline of the process used to produce successive estimates of an aggregate, and will reflect upon its implications regarding the design of the analyses of errors in provisional data. Second, particular attention will be focused upon the choice of elementary measurements of errors and suitable integrative accuracy indices, and the impact of revisions on constant price aggregates and implicit deflationary factors. Finally, the results of some empirical analysis on discordances between provisional and revised estimates will be summarily discussed on the basis of a comparison of national accounts data in Canada, Italy, and the United States.

**KEY WORDS:** Revision of data; National accounts; Accuracy indices; Constant price aggregates.

### **1. INTRODUCTION**

The evaluation of the reliability of national accounts estimates is important because of the effects of these estimates on economic policy analysis and decisions. At the same time, these estimates are difficult to evaluate, because they are affected by a very large number of complex sources of error.

Thus, it is difficult to arrive at a single reliability criterion that will be both convincing at the conceptual level and practicable. On the other hand, it seems reasonable and feasible to establish a certain number of partial criteria that could be used to measure the principal features of reliability. For a review of these various criteria, see Novak (1975) and Trivellato (1987).

The purpose of this paper is to describe and discuss critically certain recent developments in the analysis of errors in provisional national accounts estimates.

National accounts aggregates as well as many other economic aggregates and indicators are initially published in the form of provisional estimates that are subsequently subjected to various revisions.

The process of revision is determined above all by the need to ensure that information can be made quickly available and, at the same time, by the time required to collect and process all the data currently used to estimate the aggregates ("timeliness vs. accuracy," according to the effective polarization of Wilton and Smith 1974). Typically, this gives rise to a revision process characterized by a preliminary estimate and by a subsequent series of routine revisions that follow closely upon one another. From time to time, it may happen that, in order to improve the accuracy or relevance of the data, modifications are introduced into the classification and accounting plans, and/or in the basic statistical surveys and estimation methods. In turn, this leads to new evaluations, in which case, we speak of occasional or extraordinary revisions.

The characteristic feature of the errors present in provisional estimates resides in the fact that they are present in these estimates, and, by definition eliminated in the revised estimates. Thus, they can be measured by comparing the two estimates. Strictly speaking, it is true that

---

<sup>1</sup> Luigi Biggeri, Dipartimento Statistico, Università di Firenze, Italia; Ugo Trivellato, Dipartimento di Scienze Statistiche, Università di Padova, Italia.



the difference between a provisional estimate and the corresponding revised estimate is merely a measurement of the error difference between the two successive estimates. Nevertheless, when the process of revision is based on information that is progressively more complete, and on evaluation methods that are increasingly more refined, as is precisely the case in the developed countries, it is reasonable to assume that the final estimate is the closest to the true unknown value, and for purposes of comparison, to treat it as the reference value.

An understanding of the characteristics and behaviour of discordances between the provisional and revised estimates can be truly useful. In fact, significant and frequent differences are a discouraging sign respecting the quality of the basic data and/or the estimation methods used in the preliminary evaluations, and suggest at least that the results should be used carefully. (Naturally, this does not mean that, if the differences are few, this will necessarily represent a guarantee of the quality of the data. The fact that a preliminary estimate is not revised may simply mean that we do not have enough elements to improve it, regardless of its reliability). Moreover, if the differences also seem to be systematic, this may, at the same time, represent a helpful warning both to the user, who will eventually introduce correction factors; and to the producer, who may derive from this, suggestions to improve his provisional estimates.

A review of many statistical analyses of the discordance between provisional and subsequently revised estimates can be found in Trivellato (1986a). Other recent contributions have been made by Zarnowitz (1982), Mankiw, Runkle and Shapiro (1984), Mankiw and Shapiro (1986), McNees (1986), Mork (1987) and Lefrançois (1988).

In this area, there is a considerable amount of literature. Nevertheless, some questions (both methodological and substantive) concerning the way in which these analyses should be carried out have been at least partly neglected or resolved in a way that is not always satisfactory. In particular, this is true for:

- (a) The identification of a design for the analysis of the accuracy of provisional estimates that is consistent with the characteristics of the process of revision;
- (b) The definition of suitable elementary measurements of the discordances between provisional and revised estimates; the choice of suitable integrative indices that could provide adequate information on the overall accuracy of a series of provisional estimates; and, if the process of revision is carried out in several stages, of indices that could be decomposed in order to be able to verify the convergence of the provisional estimates with the final value.
- (c) The examination of the effects of errors in provisional estimates upon the derived series, especially upon constant price aggregates and implicit deflationary factors.

The three following sections will deal with these problems in order. Finally, section 5 will provide a very brief summary of the results of some empirical analyses of the discordances between provisional and revised estimates, on the basis of comparisons between national accounts data from Canada, the United States, and Italy.

## 2. A THEORETICAL OUTLINE OF THE REVISION PROCESS

A sufficiently general illustration of the process of revision of economic aggregates is shown in Table 1, which provides an immediate visualization of the relationships that exist between the reference and publication periods of the various estimates.

Three types of estimates and revisions will clearly emerge from an examination of this illustration:

Table 1  
Publication Plan for Successive Aggregate Estimates<sup>a</sup>

Publication Periods	Reference Periods															
	$t-2h+1$	$\cdot$	$t-h$	$t-h+1$	$\cdot$	$t-m$	$\cdot$	$t-1$	$t$	$t+1$	$\cdot$	$T-m$	$\cdot$	$T-1$	$T$	$T+1$
$t-2h+2$	${}_1P_{t-2h+1}$															
$\cdot$	${}_2P_{t-2h+1}$															
$\cdot$	$\cdot$															
$\cdot$	$\cdot$		${}_1P_{t-h}$													
$t-2h+m+2$	$mP_{t-2h+1}$		${}_2P_{t-h}$	${}_1P_{t-h+1}$												
$t-2h+m+3$	$r_{t-2h+1}$		$\cdot$	${}_2P_{t-h+1}$	$\cdot$											
$\cdot$	$\cdot$		$\cdot$	$\cdot$	$\cdot$	${}_1P_{t-m}$										
$\cdot$	$\cdot$		$mP_{t-h}$	$\cdot$	$\cdot$	${}_2P_{t-m}$	${}_1P_{\cdot}$									
$\cdot$	$\cdot$		$r_{t-h}$	$mP_{t-h+1}$	$\cdot$	$\cdot$	${}_2P_{\cdot}$	${}_1P_{\cdot}$								
$\cdot$	$\cdot$		$\cdot$	$r_{t-h+1}$	$\cdot$	$\cdot$	$\cdot$	${}_2P_{t-1}$	${}_1P_t$							
$\cdot$	$\cdot$		$\cdot$	$\cdot$	$\cdot$	$mP_{t-m}$	$\cdot$	${}_2P_t$	${}_1P_{t+1}$							
$t+2$	$b^T_{t-2h+1}$		$\cdot$	$b^T_{t-h}$	$\cdot$	$r_{t-m}$	$mP_{\cdot}$	$\cdot$	${}_2P_{t+1}$	$\cdot$						
$\cdot$	$\cdot$		$\cdot$	$\cdot$	$\cdot$	$r_{\cdot}$	$mP_{\cdot}$	$\cdot$	${}_2P_{t+1}$	$\cdot$						
$\cdot$	$\cdot$		$\cdot$	$\cdot$	$\cdot$	$r_{\cdot}$	$mP_{t-1}$	$\cdot$	$\cdot$	$\cdot$						
$\cdot$	$\cdot$		$\cdot$	$\cdot$	$\cdot$	$r_{t-1}$	$mP_t$	$\cdot$	$\cdot$	$\cdot$	${}_1P_{T-m}$					
$t+m+2$								$r_t$	$mP_{t+1}$	$\cdot$	${}_2P_{T-m}$	${}_1P_{\cdot}$				
$t+m+3$								$r_{t+1}$	$\cdot$	$\cdot$	${}_2P_{\cdot}$	${}_1P_{\cdot}$				
$\cdot$				$b^T_{t-h+1}$	$\cdot$	$b^T_{t-m}$	$\cdot$	$b^T_{t-1}$	$b^T_t$	$\cdot$	$\cdot$	${}_2P_{\cdot}$	${}_1P_{T-1}$			
$\cdot$											$\cdot$	${}_2P_{\cdot}$	${}_1P_{T-1}$			
$T+2$											$\cdot$	$mP_{T-m}$	${}_1P_T$			
$T+3$											$\cdot$	$r_{T-m}$	$mP_{\cdot}$	$\cdot$	${}_2P_T$	${}_1P_{T+1}$
												$r^*$	$mP^*$	$\cdot$	${}_2P^*_{T+1}$	

<sup>a</sup> The process of revision provides  $m$  successive current revisions ( ${}_2P, \dots, mP, r$ ), one benchmark revision each  $h$  period ( ${}_bP$ ), and one extraordinary revision when the retrospective series is reconstructed on the basis of time  $T + 2$ . (Source: Biggeri 1984, p.24).

- (a) those lying on the main diagonal and the lower diagonals describe a current process of revision that takes place in  $m$  steps; from the preliminary estimate  ${}_1p_t$  to the final estimate  $r_t$  (final in relation to the current process of revision);
- (b) those of the  ${}_br_t$  type, which lie horizontally, incorporate the benchmark controls where the series is reconstructed with reference to a period that goes from one benchmark to the next;
- (c) those of the  ${}_sr_t$  type are the result of an occasional revision and also lie horizontally; they concern the retrospective reconstruction of the series for a generally rather long period of time.

The figure also shows how the presence of benchmark and occasional revisions that are superimposed upon the current revision process results in mixed revisions. Thus, the revision that is carried out is not homogeneous with either the previous or the subsequent revisions carried out in the current estimation sequence.

If we have available a chronological series of provisional estimates of an aggregate for times 1 to  $n$  and that of its corresponding revised estimates, the problem of evaluating the accuracy of the first in relation to the second can be formally reduced to that of evaluating the validity of these forecasts, a problem that has been amply covered in the literature. Nevertheless, the mechanism of revising national accounts aggregates, an illustration of which is shown in Table 1, has certain characteristic features that we must keep in mind. First, it is important to remember that analyses of the process of revision explicitly take into account the existence of the three types of revisions listed above. Moreover, in order to understand correctly the characteristics of the revisions under consideration, the analyses must also exclude all comparisons that involve mixed revisions.

A second consideration of particular importance concerns the methods and criteria used to analyze the accuracy of provisional estimates. Most studies in the statistical and economic literature on the revision of national accounts aim to establish the accuracy of provisional estimates on the basis of statistical measurements that are generally descriptive. Even though very different, and interesting, approaches have been recently proposed (for example, by Mankiw and Shapiro, 1986; and by Lefrançois, 1988), they cannot be entirely satisfactory for an analysis of the properties of provisional data. In fact, they represent convenient integrations, but cannot offer a real alternative to the analysis of the properties of provisional data, for which it is important to adopt criteria and accuracy measurements that are essentially descriptive.

Finally, when carrying out an analysis of the accuracy of provisional estimates, it seems preferable to treat the provisional and final estimates in a non-symmetrical fashion, and to consider the revised series  $r_t$  as the reference series. This choice is motivated precisely by the fact that we propose to evaluate the accuracy of the provisional estimates on the basis of the final estimates (or, in any case, those that are similar to the latter in the comparison under consideration).

### 3. ELEMENTARY MEASUREMENTS AND INTEGRATIVE ERROR INDICATORS IN PROVISIONAL ESTIMATES

#### 3.1 The Choice Between Errors and Relative Errors

The error in a provisional estimate ( $p_t$ ) of the level of an economic aggregate can be obtained as follows:

$$v_t = p_t - r_t. \quad (1)$$



However, this could be inadequate to compare the accuracy of several aggregates, since the results depend upon the measuring unit and the order of magnitude of the aggregate under consideration. In this case, it may be preferable to use the relative error, which is defined as follows:

$$e_t = (p_t - r_t)/r_t = p_t/r_t - 1. \quad (2)$$

The choice between analyzing the errors or the relative errors can be a crucial one in many ways, and deserves a more in-depth discussion. As we will see in section 3.2, in fact, the pertinent integrative accuracy indices for provisional estimates are simple averages of  $v_t$  and  $e_t$  respectively (or of their suitable transformations). Thus, the use of simple means is reasonable if the series have been derived from a purely random process, particularly if they do not have a trend component. In the opposite case, we lose information, and the analysis may lead to obscure or even misleading conclusions. For this reason, it is important to carry out preliminary verifications, which can be done using various tests (see for example Malinvaud 1969, p. 473-481; Kendall 1973, p. 22-28; Box and Jenkins 1970, p. 34-36 and 287-298).

In relation to the specific problem of choosing between errors and relative errors, a particularly useful criterion is offered by the analysis of the parameters of a suitable model of the provisional and final data. In its simplest form, this can be specified as follows:

$$p_t = \alpha + \beta r_t + \epsilon_t, \quad (3)$$

where  $\epsilon_t$  is the random error. From model (3) we can obtain:

$$v_t = \alpha + (\beta - 1)r_t + \epsilon_t, \quad (4)$$

$$e_t = (\beta - 1) + \alpha \frac{1}{r_t} + \frac{\epsilon_t}{r_t}. \quad (5)$$

Formulae (4) and (5) show that, generally, both the error and the relative error of the provisional estimate depend upon the level of the corresponding final estimate. From these we can then derive the conditions that must be satisfied in order to justify the use of integrative indices based on the errors or relative errors respectively:

- (a)  $v_t$  does not depend upon  $r_t$  if  $\beta$  equals 1, and the  $\epsilon_t$ s are homoscedastic, and not temporally correlated;
- (b)  $e_t$  does not depend upon  $r_t$  if  $\alpha$  equals zero, and the variance of the  $\epsilon_t$ s is proportional to the square of the level of the series (the  $\epsilon_t$ s are not temporally correlated).

It will be immediately clear that estimating (4) with ordinary least squares is equivalent to estimating (5) with generalized least squares (with  $E(\epsilon\epsilon') = \sigma_\epsilon^2 \Omega$ , where  $\Omega$  is a diagonal matrix with  $w_{it} = r_t^2$ ). It is also immediately evident that (3) and (4) differ only by 1 in the angular coefficient. Thus, in order to choose between analyzing the errors or the relative errors, it becomes crucial to verify the homoscedasticity of  $\epsilon_t$  in (3). In order to do this, Trivellato, Di Fonzo, and Rettore (1986) developed a simple non-parametric test based on the order of the estimated residuals, which does not depend upon specifying a particular stochastic structure for equation (3); this conforms with the little we know *a priori* about the relationship that exists between provisional and final estimates.

This test can also be used to examine the hypothesis of the stability of the parameters in equation (3). In fact, it is evident that conditions (a) and (b) above are not exhaustive, and that a possible reason for their lack of validity is the presence of instabilities in the revision process. The importance of carrying out this type of verification is based, among other things, on the fact that we study the superposition of occasional revisions onto the current revision process. If the results of the test favour the hypothesis of stability, it is appropriate to analyze the current revision process as a whole. In the opposite case, the analysis must be carried out separately for the two periods preceding and following the occasional revision.

Finally, it is important to point out that specification (3) is intentionally stylized and can be essentially used to explore the process of revising an annual series, when the residual  $\epsilon_t$ s are not self-correlated. Nevertheless, it is easy to generalize it in various ways: (i) by introducing a vector of non-random explanatory variables in order to take into account any eventual factors that may affect the process of revision (benchmark revisions, seasonal factors, *etc.*); (ii) by assuming that the residuals are self-correlated; (iii) by modelling jointly the process of revision of several series using a seemingly unrelated regression equations system. For these developments, see Trivellato and Rettore (1986), and Bordignon and Trivellato (1989).

### 3.2 Integrative Indices of the Accuracy of Provisional Estimates and “Low Coherence”

In order to characterize suitable integrative indices of the overall accuracy of a provisional estimates vector, it is important to refer to the property of “low coherence,” as defined by Trivellato (1986b). In substance, in reference to two series of provisional estimates corresponding to the same series of final estimates, this requires that if the first series shows errors that are smaller than or equal to those of the second in absolute value, and if significant inequality is present in at least one case, the accuracy index of the first series must be smaller than the index of the second series, and thus indicates that the first series of provisional estimates is closer to the final data than the second.

As we will see, the central reference to integrative low coherence indices does not lead to any significant innovations in relation to the normally used measurements (we will essentially use the mean absolute error and the mean quadratic error). This concept, together with the choice of analyzing either the errors or the relative errors, and the eventual identification of homogeneous sub-periods in the revision process can nevertheless lead to an adequate implementation of empirical analyses, while avoiding the lack of accuracy that can often be found in the literature.

It is also important to consider the decompositions of the low-coherence integrative indices. From a slightly different point of view, the presence and weight of specific components can eventually be revealed by looking at the estimated parameters in equation (3) (Mincer and Zarnowitz 1969; Hatanaka 1974; Hempenius 1980; Trivellato 1986a).

The indices that we should emphasize are the following:

#### (I) Errors

If the criteria used to choose between the errors and the relative errors (see section 3.1) lead to the implementation of an analysis of errors, two low-coherence integrative indices would be the mean absolute error and the square root of the mean quadratic error:

$$\bar{v}' = \sum |v_t| / n, \quad (6)$$

$$d_v = \sqrt{(\sum v_t^2 / n)}, \quad (7)$$

where the sum is extended to  $n$  terms in the series.

The following decomposition, similar to Theil's asymmetric decomposition (1966), but developed by treating  $r_t$  as a reference series, makes it possible to show the deficiencies in the performance of provisional estimates:

$$d_v^2 = (\bar{p} - \bar{r})^2 + (s_r - \hat{\rho} s_p)^2 + (1 - \hat{\rho}^2) s_p^2, \quad (8)$$

where  $s_r$  and  $s_p$  are the standard deviations of  $r_t$  and  $p_t$  respectively, and  $\hat{\rho}$  is the coefficient of correlation between  $p_t$  and  $r_t$ . From formula (8) we can derive the following relative decomposition:

$$1 = U_v^M + U_v^R + U_v^D. \quad (9)$$

The three terms in equation (9) represent fractions of the mean quadratic error and can be interpreted as follows: a bias component between the means of the two series,  $U_v^M$ ; a regression component attributable to a deviation of 1 from the regression coefficient of  $p_t$  over  $r_t$ ,  $U_v^R$ ; and a random error component attributable to the variance of the regression errors,  $U_v^D$ .

When assessing the quality of estimates, it is also useful to take into account the mean and mean quadratic deviation of the errors; that is,  $\bar{v} = \sum v_t/n = \bar{p} - \bar{r}$  and  $s_v = \sqrt{(\sum (v_t - \bar{v})^2/n)}$  respectively. Clearly,  $\bar{v}$  is not an index of the accuracy of provisional estimates. It provides information only about the direction and dimension of the mean level error. Nevertheless, it is a statistical measurement of remarkable significance for at least two reasons: (i) if  $\bar{v} \approx 0$ , then there is no bias component ( $U_v^M = 0$ ); (ii) if  $\bar{v} \neq 0$ , it will be instructive to examine  $\bar{v}$  and  $\bar{v}'$  jointly, since  $|\bar{v}| \approx |\bar{v}'|$  if and only if the errors are always (or almost always) in the same direction. Thus, the comparison between the two indices can show the presence of a possible systematic component in the level errors of the provisional estimates.

## (II) Relative Errors

When the criterion used in section 3.1 leads to the analysis of the relative errors, two suitable integrative indices are the absolute mean relative error, and the square root of the mean quadratic relative error respectively:

$$\bar{e}' = \sum |e_t|/n, \quad (10)$$

$$d_e = \sqrt{(\sum e_t^2/n)}. \quad (11)$$

The two indices are defined if  $r_t \neq 0$  for each  $t$ , a condition that remains non-restrictive for most economic aggregates.

A reasonable decomposition of  $d_e$  is as follows:

$$d_e^2 = [(\bar{p}/\bar{r}) - 1]^2 + \frac{1}{n} \sum [(p_t/r_t) - (\bar{p}/\bar{r})]^2, \quad (12)$$

from which can be derived the relative decomposition

$$1 = U_e^b + U_e^c, \quad (13)$$

$U_e^b$  being the fraction of the mean quadratic relative error due bias, and  $U_e^c$  the fraction due to the random component.



Two other specific components of the lack of accuracy should also be mentioned; these are, the mean and mean quadratic deviation of relative errors; that is,  $\bar{e} = \sum e_t/n = (\bar{p}/\bar{r}) - 1$  and  $s_e = \sqrt{(\sum (e_t - \bar{e})^2/n)}$  respectively. *Mutatis mutandis*, what we have said for  $\bar{v}$  is also valid for  $\bar{e}$ , both in terms of its meaning and for the comparison with  $\bar{e}'$ .

### 3.3 Decomposition of the Revision Process from the Preliminary to the Final Estimate

The accuracy indices discussed until now involve a comparison between two vectors  $p$  and  $r$ . However, as we have already observed, the process of revision of economic aggregates cannot usually be completed in a single stage. Thus, how can we summarily estimate the convergence of the succession of provisional estimates with the final estimate?

In this section, we will discuss the decomposition of a process of revision that takes place in two stages (a procedure for the most common situation of  $m - 1$  steps, can be found in Wilton and Smith 1974). This decomposition will be evaluated with reference to (i) errors; and (ii) relative errors respectively.

When we analyze the errors, the decomposition of the mean quadratic error of  ${}_1p$  in relation to  $r$  in the two phases from  ${}_1p$  to  ${}_2p$  and from  ${}_2p$  to  $r$  is easy to obtain. Let  ${}_1v = {}_1p - {}_2p$ ,  ${}_2v = {}_2p - r$ ,  ${}_Tv = {}_1p - r$  and let  $d_{v1}^2$ ,  $d_{v2}^2$ ,  $d_{vT}^2$  show the mean quadratic error associated with vectors  ${}_1v$ ,  ${}_2v$  and  ${}_Tv$  respectively. The result is:

$$d_{vT}^2 = d_{v1}^2 + d_{v2}^2 + \frac{2}{n} \sum {}_1v_t {}_2v_t, \quad (14)$$

from which we derive the relative decomposition:

$$1 = D_I + D_{II} + D_{I,II}, \quad (15)$$

where the first two components represent the fraction of  $d_{vT}^2$  that can be attributed to the mean quadratic error of the first and second revision respectively, while  $D_{I,II}$  represents the interaction between  ${}_1v$  and  ${}_2v$ .

On the other hand, if we consider the relative errors, an integrative decomposition of the process of revision in the two stages may be problematic. Useful indications can be derived from the equation  $({}_1p_t - r_t)/r_t = ({}_1p_t - {}_2p_t)/r_t + ({}_2p_t - r_t)/r_t$ ; thus, by dividing the two members by  $({}_1p_t - r_t)/r_t$  and taking into account the mean values, we obtain:

$$1 = \frac{1}{n} \sum ({}_1p_t - {}_2p_t)/({}_1p_t - r_t) + \frac{1}{n} \sum ({}_2p_t - r_t)/({}_1p_t - r_t). \quad (16)$$

If we assume a stable order of estimates of the  ${}_1p_t$  type  $< {}_2p_t < r_t$  for each  $t$ , the two terms of equation (16) can be interpreted as mean fractions of the discordance between  ${}_1p$  and  $r$  eliminated by the first and second revision respectively. Nevertheless, this interpretation can be questioned in terms of order violations, especially when this occurs together with small  ${}_1p_t - r_t$  differences. In general, it seems more reasonable to discard the hypothesis that the order between estimates must be respected for the whole period under consideration. In this situation, a qualitative evaluation of the degree of stability of the order relationships between the estimates, and the importance of the two stages in the revision process can be obtained from an examination of values of  $\bar{e}$  and  $\bar{e}'$ , together with comparisons between  ${}_1p$  and  ${}_2p$ , between  ${}_2p$  and  $r$ , and between  ${}_1p$  and  $r$ . Clearly, the relationship  $|\bar{e}| = \bar{e}'$  is valid if and only if the order between the two series of estimates involved in the comparison is always respected. However, we can say that the order is "most often" respected when  $|\bar{e}| \approx \bar{e}'$  meaning that:

$$(\bar{e}' - |\bar{e}|)/\bar{e}' < 2f, \quad (17)$$

where  $f$  is a fraction of violations of a predetermined order, that is considered to be acceptable.

Equation (17) can be justified as follows: let  $e_t$  ( $t = 1, 2, \dots, n$ ); the relative errors,  $n_1$ , are strictly positive;  $n_2$ s are strictly negative; and  $n_3$ s equal zero ( $n_1 + n_2 + n_3 = n$ ). It is easy to verify the following relation:

$$\bar{e}' - |\bar{e}| = \begin{cases} 2n^{-1} \sum_{i=1}^{n_1} e_i & \text{if } \bar{e} < 0 \\ -2n^{-1} \sum_{j=1}^{n_2} e_j & \text{if } \bar{e} > 0, \end{cases}$$

where quantities  $\sum_{i=1}^{n_1} e_i$  and  $-\sum_{j=1}^{n_2} e_j$  correspond to the absolute value of the sum of the sign violations of  $\bar{e}$ .

The quantity

$$(\bar{e}' - |\bar{e}|)/\bar{e}' = \begin{cases} 2 \sum_{i=1}^{n_1} e_i / (\sum_{i=1}^{n_1} e_i - \sum_{j=1}^{n_2} e_j) & \text{if } \bar{e} < 0 \\ -2 \sum_{j=1}^{n_2} e_j / (\sum_{i=1}^{n_1} e_i - \sum_{j=1}^{n_2} e_j) & \text{if } \bar{e} > 0 \end{cases}$$

is thus an index of the importance of the sign violations of  $\bar{e}$ . This index is bounded as follows:  $0 \leq (\bar{e}' - |\bar{e}|)/\bar{e}' \leq 1$ ; the lower limit is reached when  $|\bar{e}| = \bar{e}'$ ; that is, when  $n_1 = 0$ , or when  $n_2 = 0$  (if  $n_1 = n_2 = 0$ , there are no order violation problems); and the upper limit is reached when  $\bar{e} = 0$ ; that is, when  $\sum_{i=1}^{n_1} e_i = \sum_{j=1}^{n_2} e_j$ .

The relationship  $|\bar{e}| \approx \bar{e}'$  is defined by comparing the value assumed by the index with a critical value calculated in accordance with the hypothesis of the equality, in absolute value, of all the non-zero relative errors. This produces the following result:

$$(\bar{e}' - |\bar{e}|)/\bar{e}' = \begin{cases} 2n_1/(n_1 + n_2) & \text{if } \bar{e} < 0 \\ 2n_2/(n_1 + n_2) & \text{if } \bar{e} > 0. \end{cases}$$

Under this particular hypothesis, the index is thus equal to twice the fraction of strict violations of the sign of  $\bar{e}$ .

Thus, it is possible: (i) to rank the estimates  ${}_1p$  and  ${}_2p$  on the basis of a mean criterion; that is, the values of  $\bar{e}$  associated with the various comparisons; (ii) to examine the degree of stability of the order on the basis of equation (17); (iii) to use decomposition (16), or to estimate qualitatively in any other way, the magnitude of the two stages, by observing the values of  $\bar{e}$  associated with the various comparisons, when there is a high degree of stability.

#### 4. THE IMPACT OF ERRORS IN PROVISIONAL ESTIMATES UPON CONSTANT PRICE AGGREGATES

Evidently, the revision process has certain effects upon the derived series, which often contain the information that holds the greatest interest for analysts and policy-makers. This is particularly true in the case of estimates of the rate of variation, seasonally-adjusted series, constant price aggregates, and implicit deflationary factors. The impact of revisions on measurements

of variation has been analyzed descriptively by Trivellato, Di Fonzo and Rettore (1986); and by Rao, Srinath and Quenneville (1989), who attempted to determine the best estimator of variation. The effects of revisions upon the seasonal adjustment procedures have been described in papers by Pierce (1980), Wallis (1982), and Maravall and Pierce (1983). The effects of the revision process upon constant price aggregates and implicit deflationary factors will be examined in this section.

The criteria and indices discussed in section 3.2 could evidently be used to analyze the accuracy of provisional constant price estimates. It would nevertheless be interesting to illustrate the formal characteristics of errors in these estimates, and to show the relationships that exist between estimation errors in current price and constant price aggregates, as well as the implicit deflationary factors. The emphasis will be solely upon aggregates of national economic accounts made up of the flow of goods, and the aggregates obtained using accounting balances for these goods.

Let us consider the simple case of a series of provisional estimates and the corresponding series of final estimates. By observing any aggregate at time  $t$ , consisting of  $k$  elementary goods and services, the final estimates can be obtained as follows (when we have a sum of  $k$  elementary goods and services):  $A_t = \sum p_t q_t$  is the current price aggregate;  $C_t = \sum p_o q_t$  is the constant price aggregate (0 base);  $P_t$  is the Paasche price index (that is,  $P_t = \sum p_t q_t / \sum p_o q_t$ );  $Q_t$  is the Laspeyres quantity index (that is,  $Q_t = \sum p_o q_t / \sum p_o q_o$ );  $D_t = A_t / C_t$  is the implicit deflationary factor with a Paasche structure. The corresponding provisional estimates are identified with the subscript  $p$  (thus, for example,  ${}_p A_t$  and  $A_t$  are respectively the provisional and final estimates of the current price aggregate).

In order to explain the characteristics of the errors in provisional estimates of constant price aggregates, it would be useful to remember that the evaluation of the constant price aggregate at time  $t$  may be obtained using three methods: (i) if we have a series of quantities for all the goods and services that make up the aggregate and the corresponding prices for the base year, we use the direct relationship:  $C_t = \sum p_o q_t$ ; (ii) if we have a Laspeyres-type quantity index, we multiply the value of the aggregate in the base year by this index:  $C_t = (\sum p_o q_o) Q_t$ ; (iii) if we have a Paasche-type price index, we divide the current price aggregate by this index:  $C_t = (\sum p_t q_t) / P_t$ .

In relation with the various useable evaluation criteria, the error in the provisional level estimate (as well as the relative error) of a constant price aggregate can be obtained as follows:

$${}_p C_t - C_t = \sum p_o {}_p q_t - \sum p_o q_t = \sum p_o q_o \left( \frac{\sum p_o {}_p q_t}{\sum p_o q_o} - \frac{\sum p_o q_t}{\sum p_o q_o} \right) = A_o ({}_p Q_t - Q_t), \quad (18.1)$$

$$= (\sum p_o q_o) {}_p Q_t - (\sum p_o q_o) Q_t = A_o ({}_p Q_t - Q_t), \quad (18.2)$$

$$= {}_p A_t / {}_p P_t - A_t / P_t = (\sum {}_p p_t {}_p q_t) / {}_p P_t - (\sum p_t q_t) / P_t. \quad (18.3)$$

Equations (18) show that, if we adopt the direct evaluation criterion (18.1), or extrapolation by means of an index of quantity (18.2), the estimation error coincides with the estimation error in the index of quantity, multiplied by constant  $A_o$ . On the other hand, this result cannot be obtained using the other indirect deflation criterion a price index, because, in general, (18.3) cannot be simplified in a way that is useful for our purposes. In other words, in this third case, the error in the estimate of the level of the constant price aggregate generally depends both upon errors in the provisional estimates of quantities and upon errors in the provisional estimates of prices.



The implication of this for the interpretation of errors in the estimates of constant price aggregates can be quickly made clear. Even though, theoretically, the three evaluation criteria are identical, this is far from being the case in practice, due to the availability and quantity of data. It is important to remember that evaluation at constant prices is generally obtained and a very desaggregated level, and that the aggregates are subsequently obtained by adding. These aggregates are thus evaluated with the deflation criterion, errors in the constant price estimates are not only errors in the volume estimates, but also contain errors in the price estimates.

If we now consider the relative error in the estimate of a current price aggregate and make explicit its relationship with the relative error of the corresponding constant price aggregate, we have the following equation:

$$\frac{{}_pA_t - A_t}{A_t} = \frac{{}_pC_t - C_t}{C_t} + \frac{{}_pD_t - D_t}{D_t} + \frac{{}_pC_t - C_t}{C_t} \times \frac{{}_pD_t - D_t}{D_t}.$$

This equation is evidently valid for the mean relative error of a series of  $n$  estimates, and we can obtain the following relative decomposition:

$$1 = \bar{e}(C)/\bar{e}(A) + \bar{e}(D)/\bar{e}(A) + \bar{e}(C;D)/\bar{e}(A), \quad (19)$$

where

$$\bar{e}(A) = \frac{1}{n} \sum \left( \frac{{}_pA_t - A_t}{A_t} \right)$$

( $\bar{e}(C)$  and  $\bar{e}(D)$  are defined in a similar manner, and

$$\bar{e}(C;D) = \frac{1}{n} \sum \left( \frac{{}_pC_t - C_t}{C_t} \right) \left( \frac{{}_pD_t - D_t}{D_t} \right).$$

If we leave aside the interaction component, we will then have an approximately additive decomposition: the mean relative error in the current price aggregate is equal to the sum of mean relative errors of the constant price aggregate and the implicit deflationary factor (or better still, if we take into account the estimation process: the mean relative error of the implicit deflationary factor can be approximately obtained by the difference between the mean relative errors of the current price aggregate and constant price aggregate respectively).

It is important to emphasize that the decomposition is between  $\bar{e}(C)$  and  $\bar{e}(D)$ , and not between the "error in the quantities" and the "error in the prices." We have already seen that, in general,  ${}_pC_t - C_t$  and  $\bar{e}(C)$  also reflect any errors in the provisional price estimates. On the other hand, by definition,  ${}_pD_t - D_t$ , and  $\bar{e}(D)$ , are a function of errors in the provisional estimates both of prices and of quantities. Thus, at the end, both components, "the error in the constant price estimate" and "the error in the estimate of the implicit deflationary factor" incorporate errors in the estimates of price and quantities; and the possibility of interpreting  ${}_pC_t - C_t$  and  ${}_pD_t - D_t$  as "error in the quantities" and "error in the prices" respectively can be relegated only to extreme cases.

We can identify the following extreme cases:

- (a)  ${}_pC_t - C_t$  can be interpreted as "an error in quantities" only when the evaluation of the constant price aggregate has been carried out with the direct criterion or by extrapolation using an index of quantity;

- (b)  ${}_pD_t - D_t$  can be interpreted as “an error in the prices” only in the absence of a revision in the quantities. In this case, the error in the constant price estimate is evidently zero, and the relative current price error can be expressed as a linear combination of estimation errors in the prices of  $k$  elementary goods and services:

$$({}_pA_t - A_t)/A_t = ({}_pD_t - D_t)/D_t = \sum q_t ({}_p p_t - p_t) / \sum p_t q_t = \sum a_t ({}_p p_t - p_t),$$

where

$$a_t = q_t / \sum p_t q_t.$$

- (c) If there are no revisions in the prices, we have:

$$({}_pA_t - A_t)/A_t = \sum p_t ({}_p q_t - q_t) / \sum p_t q_t = ({}_p C_t / C_t) ({}_p D_t / D_t) - 1.$$

The absence of revision in the prices is not enough to distinguish between “errors in quantity” and “errors in prices,” since the revision of quantities modifies the implicit deflationary factor.

## 5. SOME EMPIRICAL ANALYSES ON REVISIONS OF NATIONAL ACCOUNTS DATA IN CANADA, THE UNITED STATES AND ITALY

### 5.1 The Process Used to Revise National Accounts Data in the Three Countries and the Analyses Carried Out

In this last section we will summarize some of the results of empirical analyses on the revisions of national accounts in Canada, the United States and Italy.

The comparison of the characteristics of the process of revision used in the three countries contains a certain number of inevitable simplifications in the description and analyses of the revisions. For recent in-depth studies on each of the countries under consideration, refer to Lefrançois (1988) for Canada, Parker (1986) and Mork (1987) for the United States, and Di Fonzo, Rettore and Trivellato (1986) for Italy.

Charts of the processes used to revise national accounts data on a yearly and quarterly basis in the three countries under consideration are shown in Tables 2, 3 and 4.

The empirical analyses were carried out on the following data:

- (a) Canada: quarterly non-seasonally-adjusted estimates and annual current price estimates, for the period between 1953 and 1982. This is the usual database used by Statistics Canada to analyze revisions.
- (b) United States: quarterly seasonally-adjusted estimates for the period between 1968 and 1983. In this case also, this is the base normally used by the Bureau of Economic Analysis (BEA) to analyze revisions; however, contrary to the Canadian data, these do not coincide with the published data, because the BEA introduces adjustments and corrections to eliminate the effect of changing definitions.
- (c) Italy: annual estimate series, for the period between 1961 and 1985 (we did not take into account quarterly estimates, because publication was started in 1976, but was discontinued in 1982). These are data published by the *Istituto Centrale di Statistics (ISTAT)*.

By taking into account available data, it is possible to carry out comparisons between Canada and the United States for the quarterly process, and between Canada and Italy for the annual process.

**Table 2**  
Publication Plan for Quarterly and Annual Estimates of National  
Accounts Aggregates in Reference to Time *t* in Canada  
(Current Revisions Only)<sup>a</sup>

Period of Publication		Period of Reference				
Year	Month and Quarter	Quarterly Data				Yearly Data
		I	II	III	IV	
<i>t</i>	1					
	2					
	3					
	4					
	5	1 <i>P</i> <sub><i>t</i>.I</sub>				
	6					
	7					
	8	2 <i>P</i> <sub><i>t</i>.I</sub>	1 <i>P</i> <sub><i>t</i>.II</sub>			
	9					
	10					
	11	3 <i>P</i> <sub><i>t</i>.I</sub>	2 <i>P</i> <sub><i>t</i>.II</sub>	1 <i>P</i> <sub><i>t</i>.III</sub>		
	12					
<i>t</i> + 1	I					
	2	4 <i>P</i> <sub><i>t</i>.I</sub>	3 <i>P</i> <sub><i>t</i>.II</sub>	2 <i>P</i> <sub><i>t</i>.III</sub>	1 <i>P</i> <sub><i>t</i>.IV</sub>	1 <i>P</i> <sub><i>t</i></sub>
	3					
	II	5 <i>P</i> <sub><i>t</i>.I</sub>	4 <i>P</i> <sub><i>t</i>.II</sub>	3 <i>P</i> <sub><i>t</i>.III</sub>	2 <i>P</i> <sub><i>t</i>.IV</sub>	2 <i>P</i> <sub><i>t</i></sub>
	III					
	IV					
<i>t</i> + 2	I					
	II	<i>r</i> <sub><i>t</i>.I</sub>	<i>r</i> <sub><i>t</i>.II</sub>	<i>r</i> <sub><i>t</i>.III</sub>	<i>r</i> <sub><i>t</i>.IV</sub>	<i>r</i> <sub><i>t</i></sub>
	III					
	IV					

<sup>a</sup> Series shown are those used to carry out the empirical analysis on the revisions of quarterly data.

In order to carry out comparative analyses between the processes used to revise quarterly estimates in Canada and the United States, it was necessary first to establish which of the U.S. series could be used in parallel with the Canadian series. By taking into account the temporal shift between successive estimates, the available data, and the characteristics of the revision process, we chose the series in Table 3. In this way, we have, for the two countries, and for each reference quarter, three estimates that will be identified by the same letter: *P* is the first quarterly estimates series, *RT* is the chosen series of quarterly revisions; and *A1* is the first annual revisions series. To these three estimates, we added the final estimates; these coincide with the last published estimates, which we will identify with the letter *F*.

Subsequently, it seemed appropriate to carry out the following comparisons: (*P,RT*), which takes into account the effect of the quarterly revision (sub- annual). Nevertheless, it is important to point out that, in Canada, only the provisional data for the first three quarters of any year are subject to strictly quarterly revisions, since the revisions to



the fourth quarter data ( ${}_2p_{t,iv}$  in Table 2) is made up of both a quarterly and an annual revision. Among other things, this results in a very obvious difference in the behaviour of the revisions (see Lefrançois, 1988). We took this characteristic into account simply by limiting the analysis to the first three quarters of every year.

( $RT, A1$ ), which provides information on the contribution made by the annual benchmark to the quarterly process of data revision;

**Table 3**  
Publication Plan of Quarterly and Annual Estimates of National  
Accounts Aggregates in the United States<sup>a</sup>

Period of Publication		Period of Reference				Yearly Data
Year	Month and Quarter	Quarterly Data				
		I	II	III	IV	
$t$	1					
	2					
	3					
	4	${}_0P_{t.I}$				
	5	${}_1P_{t.I}$				
	6	${}_2P_{t.I}$				
	7	${}_3P_{t.I}$	${}_0P_{t.II}$			
	8		${}_1P_{t.II}$			
	9		${}_2P_{t.II}$			
	10		${}_3P_{t.II}$	${}_0P_{t.III}$		
	11			${}_1P_{t.III}$		
	12			${}_2P_{t.III}$		
			${}_3P_{t.III}$	${}_0P_{t.IV}$		
$t + 1$	1				${}_1P_{t.IV}$	
	2				${}_2P_{t.IV}$	
	3				${}_3P_{t.IV}$	
	II					
	III	${}_4P_{t.I}$	${}_4P_{t.II}$	${}_4P_{t.III}$	${}_4P_{t.IV}$	${}_1P_t$
	IV					
$t + 2$	I					
	II					
	III	${}_5P_{t.I}$	${}_5P_{t.II}$	${}_5P_{t.III}$	${}_5P_{t.IV}$	${}_2P_t$
	IV					
$t + 3$	I					
	II					
	III	$r_{t.I}$	$r_{t.II}$	$r_{t.III}$	$r_{t.IV}$	$r_t$
	IV					
$t + 5$	IV	$b r'_{t.I}$	$b r'_{t.II}$	$b r'_{t.III}$	$b r'_{t.IV}$	$b r'_t$
$t + 10$	IV	$b r''_{t.I}$	$b r''_{t.II}$	$b r''_{t.III}$	$b r''_{t.IV}$	$b r''_t$

<sup>a</sup> The series shown are those used to carry out the empirical analysis on the revisions of quarterly data.

Table 4  
Publication Plan for Quarterly and Annual Estimates of National  
Accounts Aggregates in Italy  
(Current Revisions Only)

Period of Publication		Period of Reference				
Year	Month and Quarter	Quarterly Data				Yearly Data
		I	II	III	IV	
<i>t</i>	1					
	2					
	3					
	4					
	5					
	6	$1P_{t,I}$				
	7					
	8					
	9		$1P_{t,II}$			
	10					
	11					
	12			$1P_{t,III}$		
<i>t + 1</i>	1					
	2					
	3	$2P_{t,I}$	$2P_{t,II}$	$2P_{t,III}$	$1P_{t,IV}$	$1P_t$
	II					
	III					
<i>t + 2</i>	IV					
	I	$3P_{t,I}$	$3P_{t,II}$	$3P_{t,III}$	$2P_{t,IV}$	$2P_t$
	II					
	III					
<i>t + 3</i>	IV					
	I	$r_{t,I}$	$r_{t,II}$	$r_{t,III}$	$r_{t,IV}$	$r_t$
	II					
	III					
	IV					

(A1,F), on the basis of which we were able to evaluate the effect of five-year benchmark estimates in the United States and compare them to those determined by the extraordinary revision process in Canada. In this respect, we should also point out that the comparison must be carried out with a great deal of caution, since the Canadian historical revision are mainly caused by changes in the definitions, and are not strictly comparable with the five-year data adjustment process carried out in the United States.

As far as the comparison between current revisions of annual data in Italy and Canada is concerned, we still have some minor problems, because in the two countries the data available describe a process of revision that takes place in two stages. We identified the three series of

estimates under consideration in the usual way, as follows,  ${}_1p$ ,  ${}_2p$  and  $r$ ; in this case, the interesting comparisons are:  $({}_1p, r)$ , to evaluate the overall effect of the current revision process;  $({}_2p, r)$ , to verify how much of the revision is carried out in the second stage (and indirectly also to derive information on the first stage); and the simultaneous comparison  $({}_1p, {}_2p, r)$  to evaluate the convergence of the series of estimates.

The analyses were carried out on the following four aggregates:

*PL* = Gross Domestic Product for Italy; Gross National Product for Canada and the United States.

*CF* = Final domestic consumption of households and private administrations for Italy; personal expenses for goods and services for Canada and the United States.

*CP* = Collective consumption of public administrations for Italy; Acquisition of goods and services by public administration for Canada and the United States.

*IN* = Gross Fixed Capital Formation.

It should be kept in mind that the differences between the reference accounting systems used (SNA for Canada and the United States, SEC for Italy) mean that the configuration of the aggregates is not completely homogeneous in the three countries.

The main results of the analysis of errors in estimates of the level of the aggregates are shown in Tables 5 to 8. More complete results are nevertheless available and we will refer to them as the occasion arises.

## 5.2 A Tentative Comparative Summary of the Results of the Analyses

A comparison of the three plans shown in Tables 2, 3, and 4 leads to the following conclusions.

**Table 5**  
Indices of the Accuracy of Provisional Estimates of the Level of  
Current Price Aggregates - Canada

Aggregates	$\bar{e}$	$\bar{e}'$	$s_e$	$d_e$	$U_e^b$
Comparison between <i>P</i> and <i>RT</i> (1953-83; $T = 89$ )					
<i>PL</i>	-.0018	.0034	.0042	.0046	.1567
<i>CF</i>	-.0009	.0025	.0039	.0040	.0560
<i>CP</i>	-.0041	.0106	.0193	.0197	.0425
<i>IN</i>	-.0047	.0133	.0206	.0211	.0498
Comparison between <i>RT</i> and <i>A1</i> (1953-83; $T = 69$ )					
<i>PL</i>	-.0046	.0064	.0065	.0080	.3372
<i>CF</i>	-.0054	.0069	.0086	.0102	.2877
<i>CP</i>	-.0008	.0176	.0262	.0262	.0009
<i>IN</i>	-.0075	.0205	.0270	.0280	.0721
Comparison between <i>A1</i> and <i>F</i> (1953-70; $T = 72$ )					
<i>PL</i>	-.0556	.0556	.0197	.0590	.8882
<i>CF</i>	-.0617	.0617	.0192	.0646	.9110
<i>CP</i>	-.0673	.0727	.0585	.0892	.5699
<i>IN</i>	-.0228	.0311	.0326	.0398	.0328



**Table 6**  
Indices of the Accuracy of Provisional Estimates of the  
Level of Current Price Aggregates  
Annual Data - Canada

Aggregates	$\bar{e}$	$\bar{e}'$	$s_e$	$d_e$	$U_e^b$
Comparison between ${}_1p$ and $r$ (1953-83; $n = 20$ )					
<i>PL</i>	-.0084	.0093	.0058	.0102	.6720
<i>CF</i>	-.0080	.0086	.0077	.0112	.5191
<i>CP</i>	-.0069	.0118	.0136	.0153	.2052
<i>IN</i>	-.0173	.0183	.0114	.0208	.6966
Comparison between ${}_2p$ and $r$ (1953-83; $n = 24$ )					
<i>PL</i>	-.0032	.0041	.0044	.0054	.3444
<i>CF</i>	-.0029	.0030	.0048	.0056	.2619
<i>CP</i>	-.0038	.0063	.0087	.0096	.1627
<i>IN</i>	-.0040	.0042	.0075	.0085	.2274
Comparison between $r$ and $F$ (1953-70; $n = 18$ )					
<i>PL</i>	-.0464	.0464	.0188	.0500	.8588
<i>CF</i>	-.0544	.0544	.0200	.0580	.8809
<i>CP</i>	-.0523	.0523	.0348	.0629	.6928
<i>IN</i>	-.0130	.0147	.0140	.0192	.4637

**Table 7**  
Indices of the Accuracy of Provisional Estimates of the  
Level of Current Price Aggregates - United States

Aggregates	$\bar{e}$	$\bar{e}'$	$s_e$	$d_e$	$U_e^b$
Comparison between $P$ and $RT$ (1968-83; $T = 64$ )					
<i>PL</i>	-.0017	.0031	.0040	.0044	.1459
<i>CF</i>	-.0014	.0041	.0054	.0055	.0607
<i>CP</i>	-.0013	.0050	.0064	.0065	.0433
<i>IN</i>	-.0047	.0104	.0138	.0146	.1052
Comparison between $RT$ and $A1$ (1968-83; $T = 52$ )					
<i>PL</i>	-.0054	.0064	.0071	.0089	.3706
<i>CF</i>	-.0049	.0073	.0078	.0092	.2854
<i>CP</i>	.0001	.0088	.0109	.0109	.0002
<i>IN</i>	-.0074	.0176	.0237	.0249	.0878
Comparison between $A1$ and $F$ (1968-83; $T = 52$ )					
<i>PL</i>	-.0093	.0097	.0070	.0117	.6411
<i>CF</i>	-.0040	.0070	.0088	.0096	.1662
<i>CP</i>	.0039	.0070	.0088	.0096	.1662
<i>IN</i>	-.0408	.0409	.0253	.0480	.7222

**Table 8**  
Indices of the Accuracy of Provisional Estimates of the  
Level of Current Price Aggregates  
Annual Data - Italy

Aggregates	$\bar{e}$	$\bar{e}'$	$s_e$	$d_e$	$U_e^b$
Comparison between ${}_1p$ and $r$ (1963-85; $n = 13$ )					
<i>PL</i>	-.0095	.0095	.0093	.0133	.5091
<i>CF</i>	-.0067	.0068	.0057	.0088	.5781
<i>CP</i>	-.0111	.0122	.0129	.0170	.4249
<i>IN</i>	-.0062	.0078	.0091	.0110	.3141
GR.1	-.0032	.0113	.0140	.0144	.0489
2	-.0253	.0282	.0283	.0380	.4443
3	-.0107	.0379	.0462	.0475	.0508
4	-.0133	.0208	.0271	.0302	.1937
5	-.0056	.0101	.0113	.0126	.1994
6	-.0044	.0172	.0206	.0210	.0429
7	-.0143	.0146	.0150	.0207	.4750
8	-.0061	.0132	.0170	.0181	.1121
Total	-.0106	.0106	.0106	.0150	.4987
Comparison between ${}_2p$ and $r$ (1963-85; $n = 17$ )					
<i>PL</i>	-.0030	.0037	.0054	.0062	.2299
<i>CF</i>	-.0015	.0017	.0024	.0029	.2900
<i>CP</i>	-.0013	.0049	.0063	.0064	.0426
<i>IN</i>	-.0021	.0026	.0052	.0056	.1445
GR.1	.0002	.0026	.0038	.0038	.0018
2	-.0054	.0096	.0172	.0180	.0884
3	-.0015	.0096	.0177	.0177	.0073
4	-.0058	.0060	.0105	.0120	.2304
5	.0008	.0016	.0026	.0027	.0954
6	.0011	.0055	.0085	.0085	.0167
7	-.0039	.0069	.0093	.0101	.1459
8	-.0008	.0008	.0030	.0031	.0673
Total	-.0026	.0038	.0058	.0063	.1648

Gr.1: Agriculture, food and beverage products, tobacco.

Gr.2: Textiles, clothing and shoes, skins and leather.

Gr.3: Wood and wood furniture.

Gr.4: Non-metal minerals, chemical products, extraction industries, energy and water.

Gr.5: Buildings and civil engineering.

Gr.6: Transportation and communications.

Gr.7: Other areas.

Gr.8: Building rentals.

The system used to produce and revise national accounts estimates in the United States seems to be much more structured and strongly oriented in favour of quarterly accounts; in this respect, it is important to emphasize that the preliminary estimates are published extremely fast. By the time Canada (and Italy) produce the first quarterly estimate, in the United States, this estimate has already been revised two or three times using a remarkable amount of additional direct information (Parker 1986). On the contrary, less importance is given to the annual estimate, which comes out with a delay of nine months.

On the other hand, speed seems to be a distinctive characteristic of Italian annual estimates, which come out during the first quarter of the next year; while in Canada, the first annual estimate is only available in the second quarter (the preliminary estimate published in the first quarter is a simple aggregation of quarterly data). However, this characteristic is not necessarily a good thing, and it is reasonable to think that the fast production time of the initial annual estimates can also be explained on the basis of gaps present in the Italian quarterly accounts system.

Finally, it is important to emphasize that the United States is the only country that carries out not only regular annual benchmark revisions for quarterly estimates, but also more long-term benchmark revisions based on input-output tables, and the results of wider surveys, for example, a census. This ensures greater coherence (and not only from an accounting point of view) in the estimates of the various aggregates. In Canada and in Italy there are similar revisions; however, these are only carried out occasionally and are almost always accompanied by modifications in the classification or evaluation criteria.

On the basis of Tables 5 to 8, we are able to make the following observations:

- (a) In the Canada-U.S. comparison, the pattern of current revisions of quarterly estimates is substantially homogeneous, at least as far as the aggregates in question and the limits of the revision process described by the three available series of estimates are concerned. On the average, sub-annual revisions are rather modest (in both cases the *IN* aggregate is an exception) and their behaviour is rather irregular, even though the trend is to revise upwards. The effect of the annual benchmark revision seems to be relatively more marked and systematic, especially in terms of the *PL*. There is also a tendency to correct the previous evaluation upwards. In this second stage of the process of revision, aggregate *IN* also shows higher and more irregular size errors.
- (b) In relation to annual estimates, we also found substantial behavioural similarities in the Italy-Canada comparison. On the average, the preliminary *PL* estimate underestimates the final current data by about 0.9%; this trend is very systematic and has a sufficiently stable order so that  ${}_1p \leq {}_2p \leq r$ . This leads to an underestimation of the rates of variation when they are calculated (as is generally done) on the basis of "horizontal" comparisons. We should add that the weight of the second revision is relatively more marked for Canada.

On the whole, the empirical evidence shows that the provisional estimates are valid, especially if we look at quarterly evaluations in Canada and the United States. On the other hand, the systematic character of the underestimations of preliminary annual estimates, both in Canada and in Italy, raises certain questions (even though the size of the error is certainly not significant).

This phenomenon is widespread in many countries (see Glejser and Schavey 1974). There are nevertheless some aspects of the Italian case that deserve further consideration. The first are the extreme differences in the behaviour of errors at the disaggregated level by area of economic activity, which points to situations that are certainly weak from the point of view of information: in some areas, there is a systematic underevaluation of 6-7% on the average (Marliani 1986).

The other aspect is related to the marked re-evaluation effect due to the extraordinary revisions, which generally affects all aggregates (see Di Fonzo, Rettore, and Trivellato 1986). The trend to revise the previous evaluations upwards when introducing statistical modifications in the basic statistical surveys and/or the estimation methods, is certainly not unique to Italy. The Canadian data (and indirectly, the U.S. data, through the effects of the five-year benchmark adjustments) show a similar trend.

These are signals that emphasize the weakness of national accounts systems, and lead us to consider with less optimism the small average size of the errors in provisional estimates generated on the basis of the currently used revision process.



## ACKNOWLEDGEMENTS

The results discussed in this paper were obtained within the framework of a research project financed by the Italian Ministry of Education. The data were generously provided by the Current Economic Analysis Section of Statistics Canada, for Canada; by Robert P. Parker, Assistant Director for National Accounts, BEA, for the United States; and by ISTAT, for Italy. We would like to thank Silvano Bordignon, Tomasso Di Fonzo, Gianni Marliani, and Enrico Rettore (who form part of the research team), and the two critics who, through their comments and advise, greatly improved the quality of this study. We would also like to, acknowledge the help of Cristina Pozzato and Claudio Palmieri for their help with the empirical analyses, and especially that of Gianni Marliani, who allowed us to use some of the results of his recent work (Marliani 1987). An initial version of this paper was presented at the Journées de Statistique of ASU, which were held at Lausanne from 18 to 20 May, 1987.

## REFERENCES

- BIGGERI, L. (1984). Caratteristiche e analisi dei processi di revisione nelle valutazioni di aggregati ed indici economici: alcuni confronti internazionali. *Note Economiche*, 4, 18-63.
- BORDIGNON, S., and TRIVELLATO, U. (1989). On the optimal use of provisional data in forecasting with dynamic models. *Journal of Business & Economic Statistics*, 7, 275-286.
- BOX, G.E.P., and JENKINS, G.M. (1970). *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day.
- DI FONZO, T., RETTORE, E., and TRIVELLATO, U. (1986). L'accuratezza delle stime provvisorie degli aggregati di contabilità nazionale annuale. *Errori nei dati preliminari, previsioni e politiche economiche* (Ed. U. Trivellato). Padova: Cleup, 113-145.
- GLEJSER, H., and SCHAVEY, P. (1974). An analysis of revisions on national accounts data for 40 countries. *The Review of Income and Wealth*, 20, 317-332.
- HATANAKA, M. (1974). A simple suggestion to improve the Mincer-Zarnowitz criterion for the evaluation of forecasts. *Annals of Economic and Social Measurement*, 3, 521-524.
- HEMPENIUS, A.L. (1980). *Forecast accuracy analysis applied to forecasts of the Dutch Central Planning Bureau, 1964-1978*. Katholieke Hogeschool Tilburg, Subfaculteit der Econometrie.
- KENDALL, M.G. (1973). *Time Series*. London: Griffin.
- LEFRANÇOIS, B. (1988). Application des séries chronologiques à l'étude des révisions. *The Canadian Journal of Statistics*, 16, 83-96.
- MALINVAUD, E. (1969). *Méthodes statistiques de l'économétrie*. Paris: Dunod.
- MANKIW, N.G., RUNKLE, D.E., and SHAPIRO, M.D. (1984). Are preliminary announcements of the money stock rational forecasts? *Journal of Monetary Economics*, 14, 15-27.
- MANKIW, N.G., and SHAPIRO, M.D. (1986). News or noise? An analysis of GNP revisions. *Survey of Current Business*, 66, 20-25.
- MARAVALL, A., and PIERCE, D.A. (1983). Preliminary data error and monetary aggregate targeting. *Journal of Business & Economic Statistics*, 1, 179-186.
- MARLIANI, G. (1986). L'accuratezza delle stime provvisorie del valore aggiunto per settore di attività economica. *Errori nei dati preliminari, previsioni e politiche economiche* (Ed. U. Trivellato). Padova: Cleup, 113-146.
- MARLIANI, G. (1987). Stime provvisorie e revisioni dei dati di contabilità nazionale: il caso italiano ed alcuni confronti internazionali. *Attendibilità e tempestività delle stime di contabilità nazionale* (Ed. U. Trivellato). Padova: Cleup, 35-78.

- McNEES, S.K. (1986). Estimating GNP: The trade-off between timeliness and accuracy. *New England Economic Review*, January/February, 3-10.
- MINCER, J., and ZARNOWITZ, V. (1969). The evaluation of economic forecasts. *Economic Forecasts and Expectations* (Ed. J. Mincer). New York: NBER, Columbia University Press, 3-46.
- MORK, K.A. (1987). Ain't behavin': Forecast errors and measurement errors in early GNP estimates. *Journal of Business & Economic Statistics*, 5, 165-175.
- NOVAK, G.J. (1975). Reliability criteria for national accounts. *The Review of Income and Wealth*, 21, 323-344.
- PARKER, P.R. (1986). Revisioni nelle stime iniziali del prodotto nazionale lordo trimestrale degli Stati Uniti. *Errori nei dati preliminari, previsioni e politiche economiche* (Ed. U. Trivellato). Padova: Cleup, 181-217.
- PIERCE, D.A. (1980). Data revisions with moving average seasonal adjustment procedures. *Journal of Econometrics*, 14, 95-114.
- RAO, J.N.K., SRINATH, K.P., and QUENNEVILLE, B. (1989). Estimation of level and change using current preliminary data. *Panel Surveys* (Eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh). New York: Wiley, 457-479.
- THEIL, H. (1966). *Applied Economic Forecasting*. Amsterdam: North-Holland.
- TRIVELLATO, U. (1986a) (Ed.). *Errori nei dati preliminari, previsioni e politiche economiche*. Padova: Cleup.
- TRIVELLATO, U. (1986b). Sulla valutazione dell'accuratezza di stime provvisorie di aggregati economici. *Studi in onore di Silvio Vianelli*. Palermo: Università degli Studi, 1587-1620.
- TRIVELLATO, U. (1987). Problemi e metodi di valutazione dell'attendibilità delle stime di contabilità nazionale. *Statistica*, 47, 365-388.
- TRIVELLATO, U., DI FONZO, T., and RETTORE, E. (1986). L'accuratezza delle stime provvisorie di aggregati e di indici economici: orientamenti metodologici. *Errori nei Dati Preliminari, Previsioni e Politiche Economiche* (Ed. U. Trivellato). Padova: Cleup, 61-86.
- TRIVELLATO, U., and RETTORE, E. (1986). Preliminary data errors and their impact on the forecast error of simultaneous-equation models. *Journal of Business & Economic Statistics*, 4, 445-453.
- WALLIS, K.F. (1982). Seasonal adjustment and revision of current data: Linear filters for the X-11. *Journal of the Royal Statistical Society, series A*, 145, 74-85.
- WILTON, D.A., and SMITH, P.M. (1974). The efficiency of the G.N.P. revision process: An historical analysis. Current Economic Analysis Division, Statistics Canada, Ottawa (internal paper).
- ZARNOWITZ, V. (1982). On functions, quality, and timeliness of economic information. *Journal of Business*, 55, 87-119.





# GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue (Vol. 10, No. 2 and onward) of Survey Methodology as a guide and note particularly the following points:

## 1. Layout

- 1.1 Manuscripts should be typed on white bond paper of standard size ( $8\frac{1}{2} \times 11$  inch), one side only, entirely double spaced with margins of at least  $1\frac{1}{2}$  inches on all sides.
- 1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.
- 1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.
- 1.4 Acknowledgements should appear at the end of the text.
- 1.5 Any appendix should be placed after the acknowledgements but before the list of references.

## 2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

## 3. Style

- 3.1 Avoid footnotes, abbreviations, and acronyms.
- 3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as “exp(·)” and “log(·)”, etc.
- 3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
- 3.4 Write fractions in the text using a solidus.
- 3.5 Distinguish between ambiguous characters, (e.g., w,  $\omega$ ; o, O, 0; l, 1).
- 3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

## 4. Figures and Tables

- 4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.
- 4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

## 5. References

- 5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, p. 164).
- 5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

# DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Avant de dactylographier votre texte pour le soumettre, priez d'examiner un numéro récent de Techniques d'enquête (à partir du vol. 10, n° 2) et de noter les points suivants:

- 1. **Présentation**
  - 1.1 Les textes doivent être dactylographiés sur un papier blanc de format standard (8½ par 11 pouces), sur une face seulement, à double interligne partout et avec des marges d'au moins 1½ pouce tout autour.
  - 1.2 Les textes doivent être divisés en sections numérotées portant des titres appropriés. Le nom et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.
  - 1.4 Les remerciements doivent paraître à la fin du texte.
  - 1.5 Toute annexe doit suivre les remerciements mais précéder la bibliographie.
- 2. **Résumé**
  - Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.

- 3. **Rédaction**
  - 3.1 Éviter les notes au bas des pages, les abréviations et les sigles.
  - 3.2 Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme exp(·) et log(·) etc.
  - 3.3 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin.
  - 3.4 Écrire les fractions dans le texte à l'aide d'une barre oblique.
  - 3.5 Distinguer clairement les caractères ambigus (comme w, ω; o, O; 0; 1, l).
  - 3.6 Les caractères italiques sont utilisés pour faire ressortir des mots. Indiquer ce qui doit être imprimé en italique en le soulignant dans le texte.
- 4. **Figures et tableaux**
  - 4.1 Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).
  - 4.2 Ils doivent paraître sur des pages séparées et porter une indication de l'endroit où ils doivent figurer dans le texte. (Normalement, ils doivent être insérés près du passage qui y fait référence pour la première fois.)

- 5. **Bibliographie**
  - 5.1 Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence.
  - 5.2 La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.





- TRIVELLATO, U. (1986b). Sulla valutazione dell'accuratezza di stime provvisorie di aggregati economici. *Studi in onore di Silvio Vianelli*. Palermo: Università degli Studi, 1587-1620.
- TRIVELLATO, U. (1987). Problemi e metodi di valutazione dell'attendibilità delle stime di contabilità nazionale. *Statistica*, 47, 365-388.
- TRIVELLATO, U., DI FONZO, T., et RETTORE, E. (1986). L'accuratezza delle stime provvisorie di aggregati e di indici economici: orientamenti metodologici. *Errori nei Dati Preliminari, Previsioni e Politiche Economiche* (éd. U. Trivellato). Padova: Cleup, 61-86.
- TRIVELLATO, U., et RETTORE, E. (1986). Preliminary data errors and their impact on the forecast error of simultaneous-equation models. *Journal of Business & Economic Statistics*, 4, 445-453.
- WALLIS, K.F. (1982). Seasonal adjustment and revision of current data: Linear filters for the X-11. *Journal of the Royal Statistical Society, series A*, 145, 74-85.
- WILTON, D.A., et SMITH, P.M. (1974). The efficiency of the G.N.P. revision process: An historical analysis. Division des analyses de conjoncture, Statistique Canada, Ottawa (document interne).
- ZARNOWITZ, V. (1982). On functions, quality, and timeliness of economic information. *Journal of Business*, 55, 87-119.

- BOX, G.E.P., et JENKINS, G.M. (1970). *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day.
- DI FONZO, T., RETTORE, E., et TRIVELLATO, U. (1986). L'accuratezza delle stime provvisorie degli aggregati di contabilità nazionale annuale. *Errori nei dati preliminari, previsioni e politiche economiche* (éd. U. Trivellato). Padova: Cleup, 113-145.
- GLEJSE, H., et SCHAVEY, P. (1974). An analysis of revisions on national accounts data for 40 countries. *The Review of Income and Wealth*, 20, 317-332.
- HATANAKA, M. (1974). A simple suggestion to improve the Mincer-Zarnowitz criterion for the evaluation of forecasts. *Annals of Economic and Social Measurement*, 3, 521-524.
- HEMPENIUS, A.L. (1980). *Forecast accuracy analysis applied to forecasts of the Dutch Central Planning Bureau, 1964-1978*. Katholieke Hogeschool Tilburg, Subfaculteit der Econometrie.
- KENDALL, M.G. (1973). *Time Series*. London: Griffin.
- LEFRANÇOIS, B. (1988). Application des séries chronologiques à l'étude des révisions. *La Revue Canadienne de Statistique*, 16, 83-96.
- MALINVAUD, E. (1969). *Méthodes statistiques de l'économétrie*. Paris: Dunod.
- MANKIWI, N.G., RUNKLE, D.E., et SHAPIRO, M.D. (1984). Are preliminary announcements of the money stock rational forecasts? *Journal of Monetary Economics*, 14, 15-27.
- MANKIWI, N.G., et SHAPIRO, M.D. (1986). News or noise? An analysis of GNP revisions. *Survey of Current Business*, 66, 20-25.
- MARAVALL, A., et PIERCE, D.A. (1983). Preliminary data error and monetary aggregate targeting. *Journal of Business & Economic Statistics*, 1, 179-186.
- MARLIANI, G. (1986). L'accuratezza delle stime provvisorie del valore aggiunto per settore di attività economica. *Errori nei dati preliminari, previsioni e politiche economiche* (éd. U. Trivellato). Padova: Cleup, 113-146.
- MARLIANI, G. (1987). Stime provvisorie e revisioni dei dati di contabilità nazionale: il caso italiano ed alcuni confronti internazionali. *Attendibilità e tempestività delle stime di contabilità nazionale* (éd. U. Trivellato). Padova: Cleup, 35-78.
- McNEES, S.K. (1986). Estimating GNP: The trade-off between timeliness and accuracy. *New England Economic Review*, January/February, 3-10.
- MINCER, J., et ZARNOWITZ, V. (1969). The evaluation of economic forecasts. Dans *Economic Forecasts and Expectations* (éd. J. Mincer). New York: NBER, Columbia University Press, 3-46.
- MORRIS, K.A. (1987). 'Ain't havin' it': Forecast errors and measurement errors in early GNP estimates. *Journal of Business & Economic Statistics*, 5, 165-175.
- NOVAK, G.J. (1975). Reliability criteria for national accounts. *The Review of Income and Wealth*, 21, 323-344.
- PARKER, P.R. (1986). Revisioni nelle stime iniziali del prodotto nazionale lordo trimestrale degli Stati Uniti. *Errori nei dati preliminari, previsioni e politiche economiche* (éd. U. Trivellato). Padova: Cleup, 181-217.
- PIERCE, D.A. (1980). Data revisions with moving average seasonal adjustment procedures. *Journal of Econometrics*, 14, 95-114.
- RAO, J.N.K., SRINATH, K.P., et QUENNEVILLE, B. (1989). Estimation of level and change using current preliminary data. *Panel Surveys* (éds. D. Kasprzyk, G. Duncan, G. Kalton et M.P. Singh). New York: Wiley, 457-479.
- THEIL, H. (1966). *Applied Economic Forecasting*. Amsterdam: North-Holland.
- TRIVELLATO, U. (1986a) (éd.). *Errori nei dati preliminari, previsioni e politiche economiche*. Padova: Cleup.

(b) Dans la comparaison Italie-Canada des estimations annuelles, nous remarquons aussi un comportement très uniforme. L'estimation préliminaire de  $PL$  sous-évalue en moyenne d'environ 9% la donnée finale courante; cette tendance se présente avec une systématique marquée et avec un ordre suffisamment stable selon lequel  $1^o p \leq 2^o p \leq r$ . Ceci entraîne une sous-estimation des taux de variation quand ils sont calculés, comme on le fait généralement, sur la base des comparaisons "horizontales". Il faut ajouter que le poids de la seconde révision est relativement plus marqué pour le Canada.

Dans l'ensemble, les résultats empiriques montrent que les estimations provisoires sont valides, surtout dans le cas des évaluations trimestrielles des Etats-Unis et du Canada. Au contraire, la systématique de la sous-évaluation de la part des estimations préliminaires annuelles, aussi bien au Canada qu'en Italie, laisse perplexe, bien que la dimension de l'erreur ne soit certainement pas préoccupante.

Ce phénomène est répandu dans plusieurs pays (voir Glejser et Schavey 1974). Il y a toutefois certains aspects du cas italien qui méritent plus amples réflexions. Le premier est le comportement extrêmement différencié des erreurs au niveau désagregé par branche d'activité économique, qui signale des situations où l'information est certainement faible: dans certaines branches on enregistre une sous-évaluation systématique de 6 à 7% en moyenne (Marliani 1986).

L'autre aspect est lié à l'effet marqué de réévaluation dû aux révisions extraordinaires, affectant généralement tous les agrégats (voir Di Fonzo, Rettore et Trivellato 1986). La tendance à réviser vers le haut les évaluations antérieures, à l'occasion de modifications apportées aux enquêtes statistiques de base et/ou aux méthodes d'estimation, n'est certainement pas particulière à l'Italie. Les données canadiennes elle-mêmes (et indirectement celles des Etats-Unis par l'effet de l'ajustement quinquennal aux jalons) montrent une tendance analogue. Il s'agit de signaux qui mettent l'accent sur des aspects de faiblesse des systèmes des comptes nationaux et induisent à considérer avec moins d'assurance la petite taille moyenne des erreurs des estimations provisoires générées par le processus de révision courant.

REMERCIEMENTS

Les résultats présentés dans cet article furent obtenus dans le cadre d'un projet de recherche financé par le Ministère de l'Education Nationale Italien. Les données ont été aimablement fournies par la Division de l'Analyse Economique Courante de Statistique Canada, pour le Canada, par Robert P. Parker, Directeur Adjoint pour les Comptes Nationaux auprès du BEA, pour les Etats-Unis, par l'ISTAT pour l'Italie. Nous tenons à remercier Silvano Bordignon, Tommaso Di Fonzo, Gianni Marliani et Enrico Rettore (qui faisaient partie du groupe de recherche), ainsi que les deux critiques qui, par leurs commentaires et leurs conseils, ont grandement amélioré cette étude. Nous sommes reconnaissants à Cristina Pozzato et Claudio Palmieri pour l'assistance dans les analyses empiriques, et surtout à Gianni Marliani qui nous a permis d'utiliser quelques-uns des résultats d'un de ses récents travaux (Marliani 1987). Une première version de cette étude a été présentée aux Journées de Statistique de l'ASU, qui se sont tenues à Lausanne du 18 au 20 mai 1987.

BIBLIOGRAPHIE

BIGGERI, L. (1984). Caratteristiche e analisi dei processi di revisione nelle valutazioni di aggregati ed indici economici: alcuni confronti internazionali. *Note Economiche*, 4, 18-63.

BORDIGNON, S., et TRIVELLATO, U. (1989). On the optimal use of provisional data in forecasting with dynamic models. *Journal of Business & Economic Statistics*, 7, 275-286.



Indicateurs de la précision des estimations provisoires  
du niveau des agrégats à prix courants  
Données annuelles – Italie

Tableau 8

Agrégats				
$e$	$e'$	$s_e$	$d_e$	$U_b^e$
Comparaisons entre $1P$ et $r$ (période 1963-85; $n = 13$ )				
$PL$	-.0095	.0093	.0133	.5091
$CF$	-.0067	.0068	.0057	.5781
$CP$	-.0111	.0122	.0170	.4249
$IN$	-.0062	.0078	.0091	.3141
$GR.1$	-.0032	.0113	.0140	.0489
2	-.0253	.0282	.0283	.4443
3	-.0107	.0379	.0462	.0508
4	-.0133	.0208	.0271	.1937
5	-.0056	.0101	.0113	.1994
6	-.0044	.0172	.0206	.0429
7	-.0143	.0146	.0150	.4750
8	-.0061	.0132	.0170	.1121
Total	-.0106	.0106	.0106	.4987

$PL$	-.0030	.0037	.0054	.0062	.2299
$CF$	-.0015	.0017	.0024	.0029	.2900
$CP$	-.0013	.0049	.0063	.0064	.0426
$IN$	-.0021	.0026	.0052	.0056	.1445
$GR.1$	.0002	.0026	.0038	.0038	.0018
2	-.0054	.0096	.0172	.0180	.0884
3	-.0015	.0096	.0177	.0177	.0073
4	-.0058	.0060	.0105	.0120	.2304
5	.0008	.0016	.0026	.0027	.0954
6	.0011	.0055	.0085	.0085	.0167
7	-.0039	.0069	.0093	.0101	.1459
8	-.0008	.0008	.0030	.0031	.0673
Total	-.0026	.0038	.0058	.0063	.1648

Gr.1: Agriculture; Produits alimentaires et boissons; Tabac.  
Gr.2: Textiles; Vêtement et chaussures; Peaux et cuir.  
Gr.3: Bois et meubles en bois.  
Gr.4: Minéraux non métal.; Produits chimiques; Industrie extractive; Énergie et eau.  
Gr.5: Bâtiment et génie civil.  
Gr.6: Transports et communications.  
Gr.7: Autres branches.  
Gr.8: Location de bâtiments.

(a) Dans la comparaison Canada-Etats-Unis, le profil des révisions courantes des estimations trimestrielles est substantiellement homogène, du moins en ce qui concerne les agrégats considérés et les limites du processus de révision décrit par les trois séries d'estimations disponibles. Les révisions intra-annuelles ont en moyenne des dimensions modestes (dans les deux cas l'agrégat  $IN$  fait exception) et un comportement assez irrégulier, bien que la tendance soit à réviser vers le haut. L'effet de la révision aux jalons annuels est relativement plus marqué et plus systématique, surtout en ce qui concerne  $PL$ , et tend aussi à rectifier vers le haut l'évaluation précédente; dans ce second stade du processus de révision c'est aussi l'agrégat  $IN$  qui montre des erreurs de dimension plus élevée et de nature plus irrégulière.

Nous pouvons faire les observations suivantes à partir des tableaux 5 à 8:

Tableau 6  
Indicateurs de la précision des estimations provisoires  
du niveau des agrégats à prix courants  
Données annuelles - Canada

Agrégats					$U^e_b$
$\bar{e}$	$\bar{e}'$	$s_e$	$d_e$		
Comparaisons entre $1P$ et $r$ (période 1953-83; $n = 20$ )					
$PL$	-.0084	.0093	.0058	.0102	.6720
$CF$	-.0080	.0086	.0077	.0112	.5191
$CP$	-.0069	.0118	.0136	.0153	.2052
$IN$	-.0173	.0183	.0114	.0208	.6966
Comparaisons entre $2P$ et $r$ (période 1953-83; $n = 24$ )					
$PL$	-.0032	.0041	.0044	.0054	.3444
$CF$	-.0029	.0030	.0048	.0056	.2619
$CP$	-.0038	.0063	.0087	.0096	.1627
$IN$	-.0040	.0042	.0075	.0085	.2274
Comparaisons entre $r$ et $F$ (période 1953-70; $n = 18$ )					
$PL$	-.0464	.0464	.0188	.0500	.8588
$CF$	-.0544	.0544	.0200	.0580	.8809
$CP$	-.0523	.0523	.0348	.0629	.6928
$IN$	-.0130	.0147	.0140	.0192	.4637

Tableau 7  
Indicateurs de la précision des estimations provisoires  
du niveau des agrégats à prix courants  
Données trimestrielles - États-Unis

Agrégats						$U^e_b$
$\bar{e}$	$\bar{e}'$	$s_e$	$d_e$			
Comparaisons entre $P$ et $RT$ (période 1968-83; $T = 64$ )						
$PL$	-.0017	.0031	.0040	.0044	.1459	
$CF$	-.0014	.0041	.0054	.0055	.0607	
$CP$	-.0013	.0050	.0064	.0065	.0433	
$IN$	-.0047	.0104	.0138	.0146	.1052	
Comparaisons entre $RT$ et $A1$ (période 1968-83; $T = 52$ )						
$PL$	-.0054	.0064	.0071	.0089	.3706	
$CF$	-.0049	.0073	.0078	.0092	.2854	
$CP$	.0001	.0088	.0109	.0109	.0002	
$IN$	-.0074	.0176	.0237	.0249	.0878	
Comparaisons entre $A1$ et $F$ (période 1968-83; $T = 52$ )						
$PL$	-.0093	.0097	.0070	.0117	.6411	
$CF$	-.0040	.0070	.0088	.0096	.1662	
$CP$	.0039	.0070	.0088	.0096	.1662	
$IN$	-.0408	.0409	.0253	.0480	.7222	

5.2 Une tentative de synthèse comparative des résultats des analyses

La comparaison des trois schémas des tableaux 2, 3 et 4, nous amène les réflexions suivantes.

Le système de production et de révision des estimations de comptabilité nationale des Etats-Unis apparaît beaucoup plus structuré et fortement orienté en faveur de la comptabilité trimestrielle, au sujet de laquelle il faut remarquer l'extrême rapidité dans la publication des estimations préliminaires. Lorsque le Canada produit (et l'Italie produisait) la première estimation trimestrielle, aux Etats-Unis celle-ci a déjà été révisée deux ou trois fois, en utilisant une importante quantité d'informations directes additionnelles (Parker 1986). Moins d'importance est donnée, au contraire, à l'estimation annuelle, qui sort avec un retard de neuf mois.

Par contre, la rapidité semble être un trait distinctif des estimations annuelles italiennes, puisque elles sortent au premier trimestre de l'année suivante, tandis qu'au Canada la première estimation annuelle n'est disponible qu'au second trimestre (l'estimation préliminaire publiée dans le premier trimestre est une simple agrégation des données trimestrielles). Cette caractéristique n'est cependant pas nécessairement une qualité, et il est raisonnable de penser que les temps rapides de production des premières estimations annuelles dépendent aussi des lacunes du système italien de comptabilité trimestrielle.

Enfin, il faut souligner que seuls les Etats-Unis effectuent avec régularité non seulement des révisions aux jalons annuels pour les estimations trimestrielles, mais aussi des révisions aux jalons de plus longue périodicité, fondées sur les tableaux entrée-sortie et sur les résultats de relevés plus amples, des recensements par exemple. Ceci assure une plus grande cohérence, non seulement comptable, aux estimations des différents agrégats. Au Canada et en Italie aussi il y a des révisions analogues, effectuées cependant plus occasionnellement et presque toujours accompagnées de modifications dans les critères de classification ou d'estimation.

Tableau 5

Indicateurs de la précision des estimations provisoires  
du niveau des agrégats à prix courants  
Données trimestrielles - Canada

Agrégats					$U^e_b$
Comparaisons entre P et RT (période 1953-83; T = 89)					
PL	CF	CP	IN		
-.0018	-.0009	-.0041	-.0047	.0034	.1567
.0025	.0106	.0133	.0206	.0042	.0046
.0039	.0193	.0211	.0326	.0585	.0892
.0046	.0197	.0311	.0556	.0617	.0910
.0042	.0205	.0270	.0398	.0673	.0328
Comparaisons entre RT et A1 (période 1953-83; T = 69)					
PL	CF	CP	IN		
-.0046	-.0054	-.0008	-.0075	.0064	.3372
.0069	.0176	.0262	.0270	.0086	.2877
.0064	.0262	.0280	.0326	.0590	.0009
.0065	.0311	.0398	.0556	.0617	.0721
Comparaisons entre A1 et F (période 1953-70; T = 72)					
PL	CF	CP	IN		
-.0556	-.0617	-.0673	-.0228	.0311	.8882
.0556	.0617	.0673	.0228	.0311	.5699
.0197	.0326	.0398	.0556	.0617	.9110
.0192	.0311	.0398	.0556	.0617	.0328

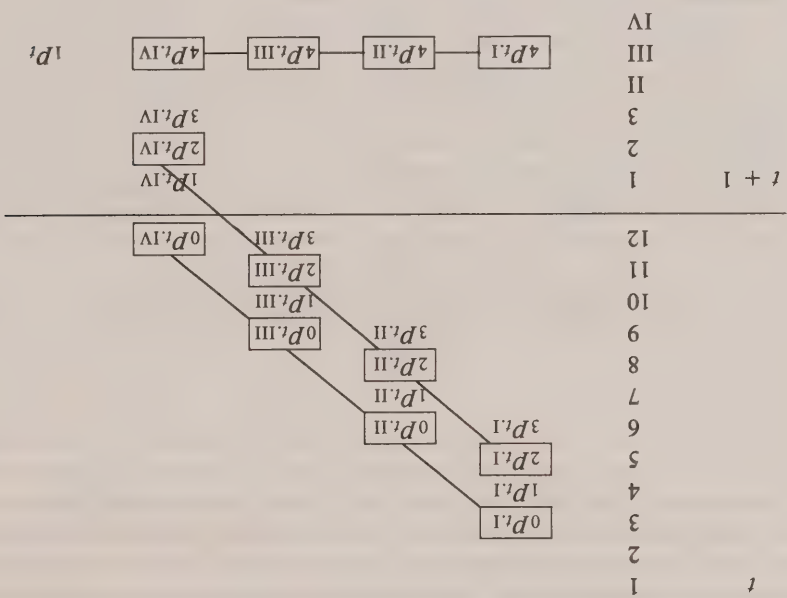


**Tableau 4**  
 Schéma de publication des estimations trimestrielles et annuelles  
 des agrégats de comptabilité nationale en Italie  
 (révisions courantes seulement)

Période de publication		Période de référence			
Année	Mois et trimestre	Données trimestrielles			Donnée annuelle
		I	II	III	
t	1				
	2				
	3				
	4				
	5				
	6	$1P_{t,1}$			
	7				
	8				
	9		$1P_{t,2}$		
	10				
	11				
	12			$1P_{t,3}$	
t + 1	1				
	2				
	3	$2P_{t,1}$			
	4				
	5				
	6		$2P_{t,2}$		
t + 2	1				
	2				
	3	$3P_{t,1}$			
	4				
	5				
	6		$3P_{t,2}$		
t + 3	1				
	2				
	3	$3P_{t,1}$			
	4				
	5				
	6		$3P_{t,2}$		
t + 4	1				
	2				
	3	$3P_{t,1}$			
	4				
	5				
	6		$3P_{t,2}$		
t + 5	1				
	2				
	3	$3P_{t,1}$			
	4				
	5				
	6		$3P_{t,2}$		
t + 6	1				
	2				
	3	$3P_{t,1}$			
	4				
	5				
	6		$3P_{t,2}$		
t + 7	1				
	2				
	3	$3P_{t,1}$			
	4				
	5				
	6		$3P_{t,2}$		
t + 8	1				
	2				
	3	$3P_{t,1}$			
	4				
	5				
	6		$3P_{t,2}$		
t + 9	1				
	2				
	3	$3P_{t,1}$			
	4				
	5				
	6		$3P_{t,2}$		
t + 10	1				
	2				
	3	$3P_{t,1}$			
	4				
	5				
	6		$3P_{t,2}$		
t + 11	1				
	2				
	3	$3P_{t,1}$			
	4				
	5				
	6		$3P_{t,2}$		
t + 12	1				
	2				
	3	$3P_{t,1}$			
	4				
	5				
	6		$3P_{t,2}$		
t + 13	1				
	2				
	3	$3P_{t,1}$			
	4				
	5				
	6		$3P_{t,2}$		
t + 14	1				
	2				
	3	$3P_{t,1}$			
	4				
	5				
	6		$3P_{t,2}$		
t + 15	1				
	2				
	3	$3P_{t,1}$			
	4				
	5				
	6		$3P_{t,2}$		
t + 16	1				
	2				
	3	$3P_{t,1}$			
	4				
	5				
	6		$3P_{t,2}$		
t + 17	1				
	2				
	3	$3P_{t,1}$			
	4				
	5				
	6		$3P_{t,2}$		
t + 18	1				
	2				
	3	$3P_{t,1}$			
	4				
	5				
	6		$3P_{t,2}$		
t + 19	1				
	2				
	3	$3P_{t,1}$			
	4				
	5				
	6		$3P_{t,2}$		
t + 20	1				
	2				
	3	$3P_{t,1}$			
	4				
	5				
	6		$3P_{t,2}$		
t + 21	1				
	2				
	3	$3P_{t,1}$			
	4				
	5				
	6		$3P_{t,2}$		
t + 22	1				
	2				
	3	$3P_{t,1}$			
	4				
	5				
	6		$3P_{t,2}$		
t + 23	1				
	2				
	3	$3P_{t,1}$			
	4				
	5				
	6		$3P_{t,2}$		
t + 24	1				
	2				
	3	$3P_{t,1}$			
	4				
	5				
	6		$3P_{t,2}$		
t + 25	1				
	2				
	3	$3P_{t,1}$			
	4				
	5				
	6		$3P_{t,2}$		
t + 26	1				
	2				
	3	$3P_{t,1}$			
	4				
	5				
	6		$3P_{t,2}$		
t + 27	1				
	2				
	3	$3P_{t,1}$			
	4				
	5				
	6		$3P_{t,2}$		
t + 28	1				
	2				
	3	$3P_{t,1}$			
	4				
	5				
	6		$3P_{t,2}$		
t + 29	1				
	2				
	3	$3P_{t,1}$			
	4				
	5				
	6		$3P_{t,2}$		
t + 30	1				
	2				
	3	$3P_{t,1}$			
	4				
	5				
	6		$3P_{t,2}$		
t + 31	1				
	2				
	3	$3P_{t,1}$			
	4				
	5				
	6		$3P_{t,2}$		
t + 32	1				
	2				
	3	$3P_{t,1}$			
	4				
	5				
	6		$3P_{t,2}$		
t + 33	1				
	2				
	3	$3P_{t,1}$			
	4				
	5				
	6		$3P_{t,2}$		
t + 34	1				
	2				
	3	$3P_{t,1}$			
	4				
	5				
	6		$3P_{t,2}$		
t + 35	1				
	2				
	3	$3P_{t,1}$			
	4				
	5				
	6		$3P_{t,2}$		
t + 36	1				
	2				
	3	$3P_{t,1}$			
	4				
	5				
	6		$3P_{t,2}$		
t + 37	1				
	2				
	3	$3P_{t,1}$			
	4				
	5				
	6		$3P_{t,2}$		
t + 38	1				
	2				
	3	$3P_{t,1}$			
	4				
	5				
	6		$3P_{t,2}$		
t + 39	1				
	2				
	3	$3P_{t,1}$			
	4				
	5				
	6		$3P_{t,2}$		
t + 40	1				
	2				
	3	$3P_{t,1}$			
	4				
	5				
	6		$3P_{t,2}$		
t + 41	1				
	2				
	3	$3P_{t,1}$			
	4				
	5				
	6		$3P_{t,2}$		
t + 42	1				
	2				
	3	$3P_{t,1}$			
	4				
	5				
	6		$3P_{t,2}$		
t + 43	1				
	2				
	3	$3P_{t,1}$			
	4				
	5				
	6		$3P_{t,2}$		
t + 44	1				
	2				
	3	$3P_{t,1}$			
	4				
	5				
	6		$3P_{t,2}$		
t + 45	1				
	2				
	3	$3P_{t,1}$			
	4				
	5				
	6		$3P_{t,2}$		
t + 46	1				
	2				
	3	$3P_{t,1}$			
	4				
	5				
	6		$3P_{t,2}$		
t + 47	1				
	2				
	3	$3P_{t,1}$			
	4				
	5				
	6		$3P_{t,2}$		
t + 48	1				
	2				
	3	$3P_{t,1}$			
	4				
	5				
	6		$3P_{t,2}$		
t + 49	1				
	2				
	3	$3P_{t,1}$			
	4				
	5				
	6		$3P_{t,2}$		
t + 50	1				
	2				
	3	$3P_{t,1}$			
	4				
	5				
	6		$3P_{t,2}$		
t + 51	1				
	2				
	3	$3P_{t,1}$			
	4				
	5				
	6		$3P_{t,2}$		
t + 52	1				
	2				
	3	$3P_{t,1}$			
	4				
	5				
	6		$3P_{t,2}$		
t + 53	1				
	2				
	3	$3P_{t,1}$			
	4				
	5				
	6		$3P_{t,2}$		
t + 54	1				
	2				
	3	$3P_{t,1}$			
	4				
	5				
	6		$3P_{t,2}$		
t + 55	1				
	2				
	3	$3P_{t,1}$			
	4				
	5				
	6		$3P_{t,2}$		
t + 56	1				
	2				
	3	$3P_{t,1}$			
	4				
	5				
	6		$3P_{t,2}$		
t + 57	1				
	2				
	3	$3P_{t,1}$			
	4				
	5				
	6		$3P_{t,2}$		
t + 58	1				
	2				
	3	$3P_{t,1}$			
	4				
	5				
	6		$3P_{t,2}$		
t + 59	1				
	2				
	3	$3P_{t,1}$			
	4				
	5				
	6		$3P_{t,2}$		
t + 60	1				
	2				
	3	$3P_{t,1}$			
	4				
	5				
	6		$3P_{t,2}$		
t + 61	1				
	2				
	3	$3P_{t,1}$			
	4				
	5				
	6		$3P_{t,2}$		
t + 62	1				
	2				
	3	$3P_{t,1}$			
	4				
	5				
	6		$3P_{t,2}$		
t + 63	1				
	2				
	3	$3P_{t,1}</$			

Schéma de publication des estimations trimestrielles et annuelles des agrégats de comptabilité nationale aux États-Unis<sup>a</sup>

Donnée annuelle	Données trimestrielles			
	I	II	III	IV
Année	Période de référence			
Mois et trimestre	Période de publication			



$t + 10$	VI	$b_{I''}^{I''}$	$b_{II''}^{I''}$	$b_{III''}^{I''}$	$b_{IV''}^{I''}$	$b_{I''}^{I''}$
$t + 5$	VI	$b_{I'}^{I'}$	$b_{II'}^{I'}$	$b_{III'}^{I'}$	$b_{IV'}^{I'}$	$b_{I'}^{I'}$
$t + 3$	VI III II I	$f_{I}$	$f_{II}$	$f_{III}$	$f_{IV}$	$f_I$
$t + 2$	VI III II I	$s_{I}^{I}$	$s_{II}^{I}$	$s_{III}^{I}$	$s_{IV}^{I}$	$s_I^{I}$

<sup>a</sup> Les séries mises en évidence sont celles que nous avons utilisées pour les analyses empiriques des révisions dans les données trimestrielles.

Les analyses ont été menées sur les quatre agrégats suivants:

$PL$  = Produit intérieur brut pour l'Italie; Produit national brut pour le Canada et les Etats-Unis.

$CF$  = Consommation finale intérieure des ménages et des administrations privées pour l'Italie; Dépenses personnelles en biens et services pour le Canada et les États-Unis.

( $A1, F$ ), qui nous permet d'évaluer l'effet de l'estimation aux jalons quinquennaux aux États-Unis et de le confronter à l'effet du processus de révision extraordinaire au Canada. À ce propos, il faut d'ailleurs observer que l'évaluation comparée doit être effectuée avec une extrême prudence, puisque les révisions historiques canadiennes sont en majeure partie causées par des changements de définitions, et ne sont donc pas strictement comparables à l'ajustement quinquennal des données aux États-Unis.

En ce qui concerne la comparaison entre les révisions courantes sur les données annuelles en Italie et au Canada, des problèmes mineurs de comparabilité subsistent, car dans les deux pays le processus de révision se déroule en deux stades. Nous désignons les trois séries d'estimations qui en résultent par  $1P$ ,  $2P$  et  $r$ . Les comparaisons d'intérêt sont alors: ( $1P, r$ ) pour évaluer l'effet global du processus de révision courant; ( $2P, r$ ) pour vérifier dans quelle mesure la révision s'accomplit au second stade (et indirectement tirer des indications sur le premier stade); la comparaison simultanée ( $1P, 2P, r$ ) pour évaluer la convergence de la succession d'estimations.

Tableau 2

Schéma de publication des estimations trimestrielles et annuelles  
des agrégats de comptabilité nationale au Canada  
(révisions courantes seulement)<sup>a</sup>

Période de publication		Période de référence				
Année	Mois et trimestre	Données trimestrielles				
		I	II	III	IV	
Donnée annuelle						
$t$	1					
	2					
	3					
	4					
	5	$1P_{t,I}$				
	6					
	7					
	8	$2P_{t,I}$				
	9					
	10					
	11	$3P_{t,I}$				
	12					
	$t + 1$	1				
		2	$4P_{t,I}$			
		3				
		4				
5						
6						
7						
8						
9						
10						
11						
12						
$t + 2$		1				
		2				
		3				
		4				
	5					
	6					
	7					
	8					
	9					
	10					
	11					
	12					
	$t + 3$	1				
		2				
		3				
		4				
5						
6						
7						
8						
9						
10						
11						
12						
$t + 4$		1				
		2				
		3				
		4				
	5					
	6					
	7					
	8					
	9					
	10					
	11					
	12					
	$t + 5$	1				
		2				
		3				
		4				
5						
6						
7						
8						
9						
10						
11						
12						
$t + 6$		1				
		2				
		3				
		4				
	5					
	6					
	7					
	8					
	9					
	10					
	11					
	12					
	$t + 7$	1				
		2				
		3				
		4				
5						
6						
7						
8						
9						
10						
11						
12						
$t + 8$		1				
		2				
		3				
		4				
	5					
	6					
	7					
	8					
	9					
	10					
	11					
	12					
	$t + 9$	1				
		2				
		3				
		4				
5						
6						
7						
8						
9						
10						
11						
12						
$t + 10$		1				
		2				
		3				
		4				
	5					
	6					
	7					
	8					
	9					
	10					
	11					
	12					
	$t + 11$	1				
		2				
		3				
		4				
5						
6						
7						
8						
9						
10						
11						
12						
$t + 12$		1				
		2				
		3				
		4				
	5					
	6					
	7					
	8					
	9					
	10					
	11					
	12					
	$t + 13$	1				
		2				
		3				
		4				
5						
6						
7						
8						
9						
10						
11						
12						
$t + 14$		1				
		2				
		3				
		4				
	5					
	6					
	7					
	8					
	9					
	10					
	11					
	12					
	$t + 15$	1				
		2				
		3				
		4				
5						
6						
7						
8						
9						
10						
11						
12						
$t + 16$		1				
		2				
		3				
		4				
	5					
	6					
	7					
	8					
	9					
	10					
	11					
	12					
	$t + 17$	1				
		2				
		3				
		4				
5						
6						
7						
8						
9						
10						
11						
12						
$t + 18$		1				
		2				
		3				
		4				
	5					
	6					
	7					
	8					
	9					
	10					
	11					
	12					
	$t + 19$	1				
		2				
		3				
		4				
5						
6						
7						
8						
9						
10						
11						
12						
$t + 20$		1				
		2				
		3				
		4				
	5					
	6					
	7					
	8					
	9					
	10					
	11					
	12					
	$t + 21$	1				
		2				
		3				
		4				
5						
6						
7						
8						
9						
10						
11						
12						
$t + 22$		1				
		2				
		3				
		4				
	5					
	6					
	7					
	8					
	9					
	10					
	11					
	12					
	$t + 23$	1				
		2				
		3				
		4				
5						
6						
7						
8						
9						
10						
11						
12						
$t + 24$		1				
		2				
		3				
		4				
	5					
	6					
	7					
	8					
	9					
	10					
	11					
	12					
	$t + 25$	1				
		2				
		3				
		4				
5						
6						
7						
8						
9						
10						
11						
12						
$t + 26$		1				
		2				
		3				
		4				
	5					
	6					
	7					
	8					
	9					
	10					
	11					
	12					
	$t + 27$	1				
		2				
		3				
		4				
5						
6						
7						
8						
9						
10						
11						
12						
$t + 28$		1				
		2				
		3				
		4				
	5					
	6					
	7					
	8					
	9					
	10					
	11					
	12					
	$t + 29$	1				
		2				
		3				
		4				
5						
6						
7						
8						
9						
10						
11						
12						
$t + 30$		1				
		2				
		3				
		4				
	5					
	6					
	7					
	8					
	9					
	10					
	11					
	12					
	$t + 31$	1				
		2				
		3				
		4				
5						
6						
7						
8						
9						
10						
11						
12						
$t + 32$		1				
		2				
		3				
		4				
	5					
	6					
	7					
	8					
	9					
	10					
	11					
	12					
	$t + 33$	1				
		2				
		3				
		4				
5						
6						
7						
8						
9						
10						
11						
12						
$t + 34$		1				
		2				
		3				
		4				
	5					
	6					
	7					
	8					
	9					
	10					
	11					
	12					
	$t + 35$	1				
		2				
		3				
		4				
5						
6						
7						
8						
9						
10						
11						
12						
$t + 36$		1				
		2				
		3				
		4				
	5					
	6					
	7					
	8					
	9					
	10					
	11</					



## 5. QUELQUES ANALYSES EMPIRIQUES SUR LES RÉVISIONS DES DONNÉES DE COMPTABILITÉ NATIONALE DU CANADA, DES ÉTATS-UNIS ET DE L'ITALIE

### 5.1 Les processus de révision des données de comptabilité nationale dans les trois pays et les analyses effectuées

Dans cette dernière section, nous présentons sommairement quelques résultats de l'analyse empirique des révisions des comptes nationaux du Canada, des États-Unis et de l'Italie. La comparaison des caractéristiques du processus de révision des trois pays considérés comporte un certain nombre de simplifications inévitables dans la description et dans l'analyse des révisions. Pour des études récentes et approfondies relatives à chacun des pays en question, il est utile de se référer à Lefrançois (1988) pour le Canada, à Parker (1986) et Mork (1987) pour les États-Unis et à Di Fonzo, Rettore et Trivellato (1986) pour l'Italie. Une schématisation des processus de révision des données de comptabilité nationale annuelle et trimestrielle dans les trois pays est reportée dans les tableaux 2, 3 et 4.

Les données sur lesquelles portent les analyses empiriques sont les suivantes:

- (a) Canada: séries des estimations trimestrielles non désaisonnalisées et des estimations annuelles, à prix courants, pour la période 1953-1982. Il s'agit de la base des données habituellement utilisée par Statistique Canada pour l'analyse des révisions.
- (b) États-Unis: séries des estimations trimestrielles désaisonnalisées, pour la période 1968-1983. Dans ce cas aussi il s'agit de la base des données habituellement utilisée par le *Bureau of Economic Analysis* (BEA) pour l'analyse des révisions, mais contrairement aux données canadiennes, elles ne coïncident pas avec les données publiées, puisque le BEA y apporte des ajustements et corrections pour éliminer l'effet de changements de définition.
- (c) Italie: séries des estimations annuelles, pour la période 1961-1985. (Nous n'avons pas considéré les estimations trimestrielles, car leur publication, commencée en 1976, s'est interrompue en 1982.) Il s'agit des données publiées par l'*Istituto Centrale di Statistica* (ISTAT).

En tenant compte des données disponibles, il est donc possible de comparer le processus trimestriel du Canada avec celui des États-Unis et le processus annuel du Canada avec celui de l'Italie. Afin d'effectuer des analyses comparatives entre les processus de révision des estimations trimestrielles du Canada et des États-Unis, il faut en premier lieu établir les séries des États-Unis pouvant être utilisées en parallèle avec les séries canadiennes. En tenant compte du décalage temporel entre les estimations successives, des données disponibles et des caractéristiques du processus de révision, nous avons choisi les séries du tableau 3. De cette façon, dans les deux pays on dispose, pour chaque trimestre de référence, de trois estimations auxquelles nous donnons le même sigle:  $P$  est la série des premières estimations trimestrielles,  $RT$  est la série des révisions trimestrielles choisie et  $A1$  est la série des premières révisions annuelles. À ces trois estimations nous ajoutons l'estimation finale, qui coïncide avec la dernière estimation publiée, que nous désignons par  $F$ .

Il apparaît alors convenable d'effectuer les comparaisons suivantes:

- ( $P, RT$ ), qui rend compte de l'effet de la révision de type trimestriel (intra-annuel). Il faut pourtant noter qu'au Canada, seules les données préliminaires des trois premiers trimestres d'une année sont sujettes à des révisions strictement trimestrielles, puisque la révision apportée aux données d'un quatrième trimestre (désignée par  $2P, 1v$  dans le tableau 2) est composée d'une révision trimestrielle et d'une révision annuelle. Ceci se traduit, entre autres, par une différence de comportement des révisions qui est nettement perceptible (voir Lefrançois 1988). On a tenu compte de cette particularité d'une façon simple, en limitant l'analyse aux seuls trois premiers trimestres de chaque année.
- ( $RT, A1$ ), qui informe de la contribution apportée par l'étalonnage annuel au processus de révision des données trimestrielles.

Cette identité est évidemment valable pour l'erreur relative moyenne d'une série de  $n$  estimations, et nous obtenons la décomposition relative suivante:

$$1 = \bar{e}(C)/\bar{e}(A) + \bar{e}(D)/\bar{e}(A) + \bar{e}(C;D)/\bar{e}(A), \quad (19)$$

où

$$\bar{e}(A) = \frac{1}{n} \sum \left( \frac{{}^p A_t - A_t}{A_t} \right)$$

( $\bar{e}(C)$  et  $\bar{e}(D)$ ) sont définis de manière analogue) et

$$\bar{e}(C;D) = \frac{1}{n} \sum \left( \frac{{}^p C_t - C_t}{C_t} \right) \left( \frac{{}^p D_t - D_t}{D_t} \right).$$

En laissant de côté la composante d'interaction, nous avons donc une décomposition approximativement additive: l'erreur relative moyenne de l'agrégat à prix courants est égale à la somme des erreurs relatives moyennes de l'agrégat à prix constants et du déflateur implicite. (Ou mieux, en ayant présent à l'esprit le processus d'estimation: l'erreur relative moyenne du déflateur implicite est approximativement égale à la différence entre les erreurs relatives moyennes de l'agrégat à prix courants et de celui à prix constants.)

La décomposition, notons-le bien, est entre  $\bar{e}(C)$  et  $\bar{e}(D)$ , et non pas entre "erreur dans les quantités" et "erreur dans les prix". Nous avons déjà vu que, en général,  ${}^p C_t - C_t$  et  $\bar{e}(C)$  reflètent aussi les erreurs dans les estimations provisoires des prix. D'autre part, par définition,  ${}^p D_t - D_t$  et  $\bar{e}(D)$  sont fonction des erreurs dans les estimations provisoires tant des prix que des quantités. En définitive, donc, les deux composantes "erreur d'estimation à prix constants" et "erreur d'estimation du déflateur implicite" incorporent toutes les deux des erreurs d'estimation de prix et quantités, et la possibilité d'interpréter  ${}^p C_t - C_t$  et  ${}^p D_t - D_t$  comme "erreur dans les quantités" et "erreur dans les prix" ne peut se faire qu'en des cas extrêmes.

Nous pouvons identifier les cas extrêmes suivants:

(a)  ${}^p C_t - C_t$  est interprétable comme "erreur dans les quantités" seulement lorsque l'évaluation de l'agrégat à prix constants se fait avec le critère direct ou par extrapolation au moyen d'un indice de quantité.

(b)  ${}^p D_t - D_t$  est interprétable comme "erreur dans les prix" seulement en l'absence de révision dans les quantités. Dans ce cas l'erreur d'estimation à prix constants est évidemment nulle, et l'erreur relative à prix courants peut s'exprimer comme une combinaison linéaire des erreurs d'estimation dans les prix des  $k$  biens et services élémentaires:

$$({}^p A_t - A_t)/A_t = ({}^p D_t - D_t)/D_t = \sum q_t ({}^p p_t - p_t)/\sum p_t q_t = \sum a_t ({}^p p_t - p_t),$$

où

$$a_t = q_t/\sum p_t q_t.$$

(c) S'il n'y a aucune révision dans les prix, on a:

$$({}^p A_t - A_t)/A_t = \sum p_t ({}^p q_t - q_t)/\sum p_t q_t = ({}^p C_t/C_t)({}^p D_t/D_t) - 1.$$

L'absence de révision dans les prix n'est donc pas suffisante pour distinguer entre "erreur dans les quantités" et "erreur dans les prix", puisque la révision dans les quantités modifie le déflateur implicite.

Considérons le cas simple d'une série d'estimations provisoires et de la série d'estimations définitives correspondante. En observant un quelconque agrégat au temps  $t$ , composé de  $k$  biens et services élémentaires, les estimations définitives s'obtiennent comme suit (la sommation se faisant sur les  $k$  biens et services élémentaires):  $A_t = \sum p_t q_t$  est l'agrégat à prix courants;  $C_t = \sum p_0 q_t$  l'agrégat à prix constants (base 0);  $P_t$  l'indice de prix, de type Paasche (c'est-à-dire,  $P_t = \sum p_t q_t / \sum p_0 q_t$ ;  $Q_t$  l'indice de quantité, de type Laspeyres (c'est-à-dire,  $Q_t = \sum p_0 q_t / \sum p_0 q_0$ );  $D_t = A_t / C_t$  le déflateur implicite, dont la structure est de type Paasche. Les estimations provisoires correspondantes sont identifiées avec l'indice  $p$  (par exemple,  ${}^p A_t$  et  $A_t$  sont respectivement l'estimation provisoire et définitive de l'agrégat à prix courants).

Pour expliciter les caractéristiques des erreurs dans les estimations provisoires des agrégats à prix constants, il est utile de rappeler que l'évaluation de l'agrégat à prix constants au temps  $t$  peut être obtenue par trois voies: (i) si nous disposons de la série des quantités pour tous les biens et services qui composent l'agrégat et des prix correspondants pour l'année de base, nous appliquons la relation directe:  $C_t = \sum p_0 q_t$ ; (ii) si nous disposons d'un indice de quantité de type Laspeyres, nous multiplions la valeur de l'agrégat dans l'année de base par cet indice:  $C_t = (\sum p_0 q_0) Q_t$ ; (iii) si nous disposons d'un indice de prix de type Paasche, nous divisons l'agrégat à prix courants par cet indice:  $C_t = (\sum p_t q_t) / P_t$ .

En relation avec les différents critères d'évaluation utilisables, l'erreur dans l'estimation provisoire du niveau (ainsi que l'erreur relative) d'un agrégat à prix constants est donnée par:

$${}^p C_t - C_t \equiv \sum p_0 {}^p q_t - \sum p_0 q_t = \sum p_0 q_0 \left( \frac{\sum p_0 {}^p q_t}{\sum p_0 q_t} - \frac{\sum p_0 q_0}{\sum p_0 q_t} \right) = A_0 ({}^p Q_t - Q_t), \quad (18.1)$$

$$\equiv (\sum p_0 q_0) {}^p Q_t - (\sum p_0 q_0) Q_t = A_0 ({}^p Q_t - Q_t), \quad (18.2)$$

$$\equiv {}^p A_t / P_t - A_t / P_t = (\sum p_t {}^p q_t) / P_t - (\sum p_t q_t) / P_t. \quad (18.3)$$

Les équations (18) montrent que si nous adoptons le critère d'évaluation directe (18.1) ou celui d'extrapolation au moyen d'un indice de quantité (18.2), l'erreur d'estimation coïncide avec l'erreur d'estimation dans l'indice de quantité, multiplié par la constante  $A_0$ . Nous n'avons pas en revanche un tel résultat en utilisant le critère indirect de déflation au moyen d'un indice de prix, parce qu'en général l'équation (18.3) ne se simplifie pas d'une façon utile. En d'autres mots, dans ce troisième cas l'erreur dans l'estimation du niveau de l'agrégat à prix constants dépend, en général, aussi bien des erreurs dans les estimations provisoires des quantités que de celles dans les prix.

L'implication pour l'interprétation des erreurs dans les estimations des agrégats à prix constants est vite tirée. Même si théoriquement les trois critères d'évaluation sont identiques, cela est loin d'être le cas en pratique, en raison de la disponibilité et de la qualité des données. Il faut se rappeler que l'évaluation à prix constants est généralement effectuée à un niveau très désagrégé, et que les agrégats sont obtenus ensuite par sommation. L'évaluation de ces agrégats se faisant donc partiellement avec le critère de la déflation, les erreurs dans les estimations à prix constants ne sont pas uniquement des erreurs dans les estimations du volume, mais incorporent aussi les erreurs dans les estimations des prix.

Si nous considérons maintenant l'erreur relative d'estimation d'un agrégat à prix courants et si nous en explicitons la relation avec l'erreur relative de l'agrégat à prix constants correspondant, nous avons l'identité suivante:

$$\frac{{}^p A_t - A_t}{{}^p A_t - A_t} = \frac{{}^p C_t - C_t}{{}^p C_t - C_t} + \frac{{}^p D_t - D_t}{{}^p D_t - D_t} + \frac{{}^p C_t}{{}^p C_t - C_t} \times \frac{{}^p D_t}{{}^p D_t - D_t}.$$



La quantité

$$\left\{ \begin{array}{l} (e' - |\bar{e}|)/\bar{e}' = 2 \sum_{i=1}^{n_1} e_i / (\sum_{i=1}^{n_1} e_i - \sum_{j=1}^{n_2} e_j) \quad \text{si } \bar{e} < 0 \\ - 2 \sum_{j=1}^{n_2} e_j / (\sum_{i=1}^{n_1} e_i - \sum_{j=1}^{n_2} e_j) \quad \text{si } \bar{e} > 0 \end{array} \right.$$

est donc un indice de l'importance des violations du signe de  $\bar{e}$ . Cet indice est borné, lorsque  $n_1 = 0$  ou  $n_2 = 0$  (si  $n_1 = n_2 = 0$  il n'existe pas de problèmes de violations de l'ordre), et la borne supérieure étant atteinte lorsque  $\bar{e} = 0$ , c'est-à-dire lorsque  $\sum_{i=1}^{n_1} e_i = \sum_{j=1}^{n_2} e_j$ . La relation  $|\bar{e}| \approx \bar{e}'$  est définie en comparant la valeur assumée par l'indice avec une valeur critique calculée sous l'hypothèse de l'égalité, en valeur absolue, de toutes les erreurs relatives non-nulles. Ceci donne le résultat suivant:

$$(e' - |\bar{e}|)/\bar{e}' = \left\{ \begin{array}{l} 2n_1/(n_1 + n_2) \quad \text{si } \bar{e} < 0 \\ 2n_2/(n_1 + n_2) \quad \text{si } \bar{e} > 0. \end{array} \right.$$

Sous cette hypothèse particulière, l'indice est égal au double de la fraction des violations strictes au signe de  $\bar{e}$ . Il est donc possible de: (i) ordonner les estimations  ${}^1p$  et  ${}^2p$  sur la base d'un critère moyen, c'est-à-dire des valeurs de  $\bar{e}$  associées aux différentes comparaisons; (ii) examiner le degré de stabilité de l'ordre sur la base de la relation (17); (iii) lorsqu'on constate un degré élevé de stabilité, utiliser la décomposition (16) ou de quelque manière estimer qualitativement l'importance des deux stades en observant les valeurs de  $\bar{e}$  associées aux différentes comparaisons.

#### 4. IMPLICATIONS DES ERREURS DANS LES ESTIMATIONS PROVISOIRES SUR LES AGRÉGATS À PRIX CONSTANTS

Le processus de révision a évidemment des implications sur les séries dérivées, qui constituent souvent l'information représentant le plus grand intérêt pour analystes et décideurs. Ceci vaut en particulier pour l'estimation du taux de variation, les séries désaisonnalisées et les agrégats à prix constants et les déflateurs implicites. L'impact des révisions sur la mesure de la variation a été analysé d'une façon descriptive par Trivellato, Di Fonzo et Rettore (1986), et par Rao, Srinath et Quenneville (1989) dans la perspective de déterminer le meilleur estimateur de la variation. Les effets des révisions sur les procédures de désaisonnalisation sont traités dans les articles de Pierce (1980), Wallis (1982) et Maravall et Pierce (1983). Les implications du processus de révision sur les agrégats à prix constants et les déflateurs implicites sont examinées dans cette section. Les critères et les indicateurs présentés dans la section 3.2 peuvent évidemment être utilisés pour l'analyse de la précision des estimations provisoires à prix constants. Il est toutefois intéressant d'illustrer les caractéristiques formelles des erreurs de ces estimations, et mettre en évidence les relations entre les erreurs d'estimation dans les agrégats à prix courants et ceux à prix constants ainsi que les déflateurs implicites. L'attention est circonscrite aux agrégats des comptes économiques nationaux constitués par des flux de marchandises, et aux agrégats obtenus par solde comptable de ceux-ci.

Quand nous analysons les erreurs, la décomposition de l'erreur quadratique moyenne de  ${}^1P$  par rapport à  $r$  dans les deux phases de  ${}^1P$  à  ${}^2P$  et de  ${}^2P$  à  $r$  est simple. Soit  ${}^1v = {}^1P - {}^2P$ ,  ${}^2v = {}^2P - r$ ,  ${}^{TV} = {}^1P - r$  et l'on indique par  $d_{v1}^{2v}$ ,  $d_{v2}^{2v}$ ,  $d_{vT}^{2v}$  l'erreur quadratique moyenne associée respectivement aux vecteurs  ${}^1v$ ,  ${}^2v$  et  ${}^{TV}$ . Il en résulte:

$$d_{vT}^{2v} = d_{v1}^{2v} + d_{v2}^{2v} + \frac{n}{2} \sum {}^1v {}^2v, \quad (14)$$

dont on tire la décomposition relative:

$$1 = D_I + D_{II} + D_{I,II}, \quad (15)$$

où les deux premières composantes représentent la fraction de  $d_{vT}^{2v}$  attribuable à l'erreur quadratique moyenne de la première et de la seconde révision respectivement, tandis que  $D_{I,II}$  représente l'interaction entre  ${}^1v$  et  ${}^2v$ .

En opérant sur les erreurs relatives, une décomposition synthétique du processus de révision dans les deux stades peut, au contraire, être problématique. D'utiles indications peuvent être tirées de l'identité  $({}^1P - r)/r = ({}^1P - {}^2P)/r + ({}^2P - r)/r$ , dont en divisant les deux membres par  $({}^1P - r)/r$ , et en considérant les valeurs moyennes, nous obtenons:

$$1 = \frac{1}{I} \sum ({}^1P - {}^2P) / ({}^1P - r) + \frac{n}{I} \sum ({}^2P - r) / ({}^1P - r). \quad (16)$$

Sous l'hypothèse d'ordre stable des estimations, du type  ${}^1P_t < {}^2P_t < r_t$  pour chaque  $t$ , les deux termes de l'identité (16) sont interprétables comme les fractions moyennes de la discordance entre  ${}^1P$  et  $r$  éliminées par la première et par la seconde révision respectivement. Cette interprétation est toutefois discutable lorsqu'il y a violation de l'ordre, surtout lorsqu'elle apparaît conjointement avec des petites différences  ${}^1P_t - r_t$ . En général, il est plus raisonnable d'écarter l'hypothèse que l'ordre entre les estimations soit respecté durant toute la période considérée. Dans cette situation, une appréciation qualitative du degré de stabilité de la relation d'ordre entre les estimations, et de l'importance des deux stades du processus de révision, peut être fournie par l'inspection des valeurs de  $\bar{e}$  et  $\bar{e}'$  associées aux comparaisons entre  ${}^1P$  et  ${}^2P$ , entre  ${}^2P$  et  $r$ . Evidemment, la relation  $|\bar{e}| = \bar{e}'$  n'est valable que si l'ordre entre les deux séries d'estimations faisant l'objet de la comparaison est toujours respecté. Nous pouvons cependant retenir que l'ordre est "le plus souvent" respecté quand  $|\bar{e}| \approx \bar{e}'$ , en entendant par là:

$$(\bar{e}' - |\bar{e}|) / \bar{e}' > 2f, \quad (17)$$

où  $f$  est une fraction de violations de l'ordre fixée à l'avance, considérée comme étant acceptable. L'équation (17) se justifie comme suit. Soient  $e_i$  ( $i = 1, 2, \dots, n$ ) les erreurs relatives,  $n_1$  étant strictement positives,  $n_2$  strictement négatives et  $n_3$  égales à zéro ( $n_1 + n_2 + n_3 = n$ ). Il est facile de vérifier la relation suivante:

$$\left\{ \begin{array}{ll} 2n^{-1} \sum_{i=1}^{n_1} e_i & \text{si } \bar{e} > 0 \\ -2n^{-1} \sum_{i=1}^{n_2} e_i & \text{si } \bar{e} < 0, \end{array} \right.$$

où les quantités  $\sum_{i=1}^{n_1} e_i$  et  $-\sum_{i=1}^{n_2} e_i$  correspondent à la valeur absolue de la somme des violations au signe de  $\bar{e}$ .

Les trois termes de l'équation (9) représentent des fractions de l'erreur quadratique moyenne interprétables comme suit: composante de biais entre les moyennes des deux séries,  $U_M^v$ ; composante de régression, attribuable à l'écartement de 1 du coefficient de la régression de  $p_i$  sur  $r_i$ ,  $U_R^v$ ; composante d'erreur aléatoire, attribuable à la variance des erreurs de régression,  $U_D^v$ . En jugeant de la qualité des estimations, il est aussi utile de considérer la moyenne et l'écart quadratique moyen des erreurs, c'est-à-dire respectivement  $\bar{v} = \sum v_i/n = \bar{p} - \bar{r}$  et  $s_v = \sqrt{(\sum (v_i - \bar{v})^2/n)}$ . Bien entendu,  $v$  n'est pas un indice de précision des estimations provisoires. Celui-ci nous informe seulement de la direction et de la dimension de l'erreur moyenne de niveau. Cette statistique est, cependant, remarquable pour au moins deux raisons: (i) si  $\bar{v} \approx 0$ , il n'y a pas de composante de biais ( $U_M^v = 0$ ); (ii) si  $\bar{v} \neq 0$ , il est alors instructif d'examiner conjointement  $\bar{v}$  et  $v$ , puisque  $|v| \approx v'$  si, et seulement si, les erreurs sont presque toujours dans la même direction. La comparaison des deux indices peut donc mettre en évidence une éventuelle composante systématique dans les erreurs du niveau des estimations provisoires.

(II) Erreurs relatives

Quand le critère de choix de la section 3.1 nous porte à l'analyse des erreurs relatives, deux indices synthétiques appropriés sont l'erreur relative absolue moyenne et la racine carrée de l'erreur relative quadratique moyenne, respectivement:

(10) 
$$e' = \sum |e_i|/n,$$

(11) 
$$d_e = \sqrt{(\sum e_i^2/n)}.$$

Les deux indices sont définis si  $r_i \neq 0$  pour chaque  $i$ , condition qui n'est pas restrictive pour la plupart des agrégats économiques.

Une décomposition raisonnable de  $d_e$  est la suivante:

(12) 
$$d_e^2 = [(\bar{p}/\bar{r}) - 1]^2 + \frac{1}{n} \sum [(p_i/r_i) - (\bar{p}/\bar{r})]^2,$$

dont découle la décomposition relative

(13) 
$$1 = U_e^p + U_e^r,$$

$U_e^p$  étant la fraction de l'erreur relative quadratique moyenne due au biais et  $U_e^r$  étant la fraction due à la composante aléatoire.

Deux autres composantes spécifiques de manque de précision, la moyenne et l'écart quadratique

moyen des erreurs relatives (c'est-à-dire respectivement  $\bar{e} = \sum e_i/n = (\bar{p}/\bar{r}) - 1$  et  $s_e = \sqrt{(\sum (e_i - \bar{e})^2/n)}$ ) sont aussi d'intérêt. *Mutatis mutandis*, ce qui a été dit pour  $v$  est valable pour  $e$ , autant pour son interprétation que pour la comparaison avec  $e'$ .

3.3 Décomposition du processus de révision de l'estimation préliminaire à la définitive

Les indicateurs de précision qui ont été présentés jusqu'ici concernent la comparaison entre deux vecteurs  $p$  et  $r$ . Nous avons déjà observé, cependant, que le processus de révision des agrégats économiques ne se termine usuellement pas après un seul stade. Comment, donc, estimer synthétiquement la convergence de la succession des estimations provisoires à l'estimation définitive?

Nous considérons ici la décomposition d'un processus de révision qui s'articule en deux stades. (Une procédure pour la situation plus générale de  $m - 1$  étapes se trouve dans Wilton et Smith 1974.) Cette décomposition sera évaluée, dans l'ordre, pour les erreurs et pour les erreurs relatives.



Il faut noter, enfin, que la spécification (3) est intentionnellement stylisée, et se prête essentiellement à l'exploration du processus de révision d'une série annuelle quand les résidus  $\epsilon_t$  ne présentent pas d'autocorrélation. Il est cependant facile de la généraliser de diverses façons: (i) en introduisant un vecteur de variables explicatives non aléatoires, pour tenir compte d'éventuels facteurs influant le processus de révision (révisions aux jalons, facteurs saisonniers, etc.); (ii) en admettant une autocorrélation dans les résidus; (iii) en modélisant conjointement le processus de révision de plusieurs séries, au moyen d'un système de régressions apparemment indépendantes. Pour ces développements, voir Trivellato et Rettore (1986) et Bordignon et Trivellato (1989).

### 3.2 Indices synthétiques de la précision des estimations provisoires et "cohérence faible"

Pour caractériser les indices synthétiques appropriés de la précision globale d'un vecteur d'estimations provisoires, il convient de se référer à la propriété de "cohérence faible" définie dans Trivellato (1986b). En substance, par référence à deux séries d'estimations provisoires correspondant à la même série d'estimations définitives, celle-ci requiert que lorsque la première série montre des erreurs en valeur absolue inférieures ou égales à celles de la seconde, l'indice de précision de la première série doit être plus petit que l'indice de la seconde série, signalant ainsi que la première série d'estimations provisoires est plus proche des données définitives que la seconde.

Comme on le verra, la référence centrale aux indices synthétiques faiblement cohérents n'innove pas dans les mesures habituellement utilisées (nous utiliserons essentiellement l'erreur absolue moyenne et l'erreur quadratique moyenne). Ce concept, uni au choix d'analyser les erreurs ou bien les erreurs relatives et à l'éventuel repérage de sous-périodes homogènes du processus de révision, permet toutefois de mener de manière appropriée les analyses empiriques, tout en évitant les fréquentes imprécisions de la littérature.

#### (1) Erreurs

Si le critère de choix entre les erreurs et les erreurs relatives (voir section 3.1) porte à mener l'analyse sur les erreurs, deux indices synthétiques faiblement cohérents sont l'erreur absolue moyenne et la racine carrée de l'erreur quadratique moyenne:

$$v' = \sum |v_i| / n, \quad (6)$$

$$d_v = \sqrt{(\sum v_i^2 / n)}, \quad (7)$$

où la sommation est étendue aux  $n$  termes de la série. La décomposition suivante, analogue à la décomposition asymétrique de Theil (1966) mais développée en traitant  $r_i$  comme série de référence, permet de mettre en évidence les insuffisances dans la performance des estimations provisoires:

$$d_v^2 = (d - r)^2 + (s_r - \rho s_d)^2 + (1 - \rho^2) s_d^2, \quad (8)$$

où  $s_r$  et  $s_d$  sont les écarts-typé respectifs de  $r_i$  et  $d_i$ , et  $\rho$  est le coefficient de corrélation entre  $d_i$  et  $r_i$ . De la relation (8) on tire la décomposition relative suivante:

$$1 = U_v^M + U_v^R + U_v^D. \quad (9)$$

synthétiques pertinents de précision des estimations provisoires sont des moyennes simples respectivement des  $v_i$  et des  $e_i$  (ou de leurs transformations convenables). Or, l'emploi de moyennes simples est raisonnable si les séries proviennent d'un processus purement aléatoire, et en particulier si elles ne contiennent pas une composante de tendance. Dans le cas contraire, de l'information est perdue et l'analyse peut être peu éclairante et, à la limite, fourvoyante. Pour cette raison des vérifications préalables s'imposent, qui peuvent être menées avec plusieurs tests (voir, par exemple, Malinvaud 1969, p. 473-481; Kendall 1973, p. 22-28; Box et Jenkins 1970, p. 34-36 et 287-298).

Par rapport au problème spécifique du choix entre erreurs et erreurs relatives, un critère particulièrement utile est offert par l'analyse des paramètres d'un modèle convenable entre les données provisoires et les données définitives. Dans sa formulation la plus simple, celui-ci peut être spécifié comme :

$$(3) \quad p_i = \alpha + \beta r_i + e_i,$$

où  $e_i$  est l'erreur aléatoire. Du modèle (3) il est immédiat de déduire :

$$(4) \quad v_i = \alpha + (\beta - 1)r_i + e_i,$$

$$(5) \quad e_i = (\beta - 1) + \alpha \frac{1}{r_i} + \frac{r_i}{e_i}.$$

Les relations (4) et (5) mettent en évidence que, en général, aussi bien l'erreur que l'erreur relative de l'estimation provisoire dépendent du niveau de l'estimation définitive correspondante. Nous pouvons aussi en déduire les conditions qui doivent être satisfaites pour que l'emploi d'indices synthétiques basés respectivement sur les erreurs ou bien sur les erreurs relatives soit justifié :

(a)  $v_i$  ne dépend pas de  $r_i$  si  $\beta$  est égal à un et les  $e_i$  sont homoscédastiques (et non-corrélés temporellement);

(b)  $e_i$  ne dépend pas de  $r_i$  si  $\alpha$  est égal à zéro, et si la variance des  $e_i$  est proportionnelle au carré du niveau de la série (les  $e_i$  étant non-corrélés temporellement).

On observe immédiatement (4) avec les moindres carrés ordinaires équivalant à estimer l'équation (5) avec les moindres carrés généralisés (avec  $E(e_i^2) = \sigma_i^2 \Omega$ , où  $\Omega$  est une matrice diagonale avec  $w_{ii} = r_i^2$ ). Il est aussi immédiat de noter que les équations (3) et (4) ne diffèrent que de 1 dans le coefficient angulaire. La vérification de l'homoscédastité des  $e_i$  dans l'équation (3) devient donc cruciale pour le choix entre l'analyse des erreurs ou des erreurs relatives. Dans ce but, Trivellato, Di Fonzo et Rettore (1986) ont élaboré un simple test non-paramétrique, basé sur l'ordre des résidus estimés, qui n'engage pas à spécifier une structure stochastique particulière pour l'équation (3), ce qui est en accord avec le peu de connaissance que nous avons a priori sur la relation entre les estimations provisoires et définitives.

Ce test peut aussi servir à l'examen de l'hypothèse de stabilité des paramètres de l'équation (3). Il est en effet évident que les conditions (a) et (b) considérées plus haut ne sont pas exhaustives, et qu'une raison plausible pour qu'elles ne soient pas satisfaites est la présence d'instabilités dans le processus de révision. L'intérêt de vérifications de ce type est, entre autres, d'étudier la superposition d'une révision occasionnelle au processus de révision courant. Si le résultat du test favorise l'hypothèse de stabilité, il est approprié d'analyser le processus de révision courant dans son entier. Dans le cas contraire, l'analyse doit être menée séparément dans les deux périodes précédente et successive à la révision occasionnelle.

(c) Celles qui sont la conséquence d'une révision occasionnelle ( $r_i$ ), elles aussi placées horizontalement, qui concernent la reconstruction rétrospective de la série pour une période généralement assez longue.

Ce schéma met aussi en évidence que la présence de révisions aux jalons et occasionnelles, qui se superposent au processus de révision courant, résulte en des révisions mixtes. La révision que l'on effectue ainsi n'est homogène ni avec les précédentes, ni avec les suivantes dans la séquence des estimations courantes.

En disposant d'une série chronologique d'estimations provisoires d'un agrégat pour les temps de 1 à  $n$  et de celle des estimations révisées correspondantes, le problème de l'évaluation de la précision des premières par rapport aux deuxièmes est reconductible, sur le plan formel, à celui de l'évaluation de la validité des prévisions, sur lequel existe une ample littérature. Toutefois le mécanisme de révision des agrégats de comptabilité nationale, schématisé dans le tableau 1, présente certaines particularités qu'il faut avoir présent à l'esprit.

Tout d'abord, il est opportun que les analyses du processus de révision tiennent compte explicitement de l'existence des trois types de révision mentionnés plus haut. De plus, pour saisir correctement les caractéristiques des révisions en question, toutes les comparaisons qui entraînent des révisions mixtes doivent être exclues de l'analyse.

Une deuxième considération particulièrement importante concerne les méthodes et critères pour l'analyse de la précision des estimations provisoires. La plupart des travaux dans la littérature statistique et économique portant sur les révisions des comptes nationaux, visent à établir la précision des estimations provisoires par le moyen de mesures statistiques, généralement descriptives. Bien que différentes approches, fort intéressantes, furent récemment proposées (par exemple, par Mankiw et Shapiro 1986, et par Lefrançois 1988), elles ne sont peut-être pas entièrement satisfaisantes pour une analyse des propriétés des données provisoires. En effet, elles constituent des intégrations convenables mais non des alternatives à l'analyse de base des propriétés des données provisoires, pour laquelle il convient d'adopter des critères et des mesures de précision essentiellement descriptifs.

Enfin, en effectuant l'analyse de la précision des estimations provisoires, il semble approprié de traiter de façon non symétrique les estimations provisoires et les définitions et de considérer la série révisée  $r_i$  comme série de référence. Ce choix est motivé précisément du fait que nous nous proposons d'évaluer la précision des estimations provisoires, étant données les estimations définitives (ou de toutes façons assimilables à celles-ci dans la comparaison en question).

### 3. MESURES ÉLÉMENTAIRES ET INDICATEURS SYNTHÉTIQUES DES ERREURS DANS LES ESTIMATIONS PROVISOIRES

#### 3.1 Sur le choix entre erreurs et erreurs relatives

L'erreur dans l'estimation provisoire ( $p_i$ ) du niveau d'un agrégat économique est donnée par

$$(1) \quad v_i = p_i - r_i$$

Celle-ci peut être inappropriée pour comparer la précision de plusieurs agrégats, parce que les résultats dépendent de l'unité de mesure et de l'ordre de grandeur de l'agrégat considéré. Il peut alors être préférable d'utiliser l'erreur relative, définie comme

$$(2) \quad e_i = (p_i - r_i) / r_i = p_i / r_i - 1.$$

Le choix entre l'analyse des erreurs ou des erreurs relatives est sous des nombreux aspects crucial, et mérite d'être approfondi. En effet, comme on le verra dans la section 3.2, les indices



Tableau 1  
Schéma de publication des estimations successives d'un agrégat<sup>a</sup>

Période de publication	Période de référence									
	$t-2h+1$	$\cdot$	$t-h$	$t-h+1$	$\cdot$	$t-m$	$\cdot$	$t-1$	$t$	$t+1$
$t-2h+2$	$1^P_{t-2h+1}$ $2^P_{t-2h+1}$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$
$\cdot$	$\cdot$	$\cdot$	$1^P_{t-h}$	$\cdot$	$\cdot$	$1^P_{t-m}$	$1^P_{\cdot}$	$1^P_{\cdot}$	$1^P_{t-1}$	$1^P_{t+1}$
$t-2h+m+2$	$m^P_{t-2h+1}$	$\cdot$	$2^P_{t-h}$	$1^P_{t-h+1}$	$\cdot$	$2^P_{t-m}$	$1^P_{\cdot}$	$1^P_{t-1}$	$1^P_{t+1}$	$1^P_{T-m}$
$t-2h+m+3$	$t_{t-2h+1}$	$\cdot$	$\cdot$	$2^P_{t-h+1}$	$\cdot$	$\cdot$	$2^P_{\cdot}$	$2^P_{t-1}$	$2^P_{t+1}$	$2^P_{T-m}$
$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$
$\cdot$	$\cdot$	$\cdot$	$m^P_{t-h}$	$\cdot$	$\cdot$	$2^P_{t-m}$	$1^P_{\cdot}$	$2^P_{t-1}$	$2^P_{t+1}$	$2^P_{T-m}$
$\cdot$	$\cdot$	$\cdot$	$t_{t-h}$	$m^P_{t-h+1}$	$\cdot$	$\cdot$	$2^P_{\cdot}$	$2^P_{t-1}$	$2^P_{t+1}$	$2^P_{T-m}$
$\cdot$	$\cdot$	$\cdot$	$t_{t-h+1}$	$\cdot$	$\cdot$	$\cdot$	$2^P_{\cdot}$	$2^P_{t-1}$	$2^P_{t+1}$	$2^P_{T-m}$
$t+2$	$b^P_{t-2h+1}$	$\cdot$	$b^P_{t-h}$	$\cdot$	$m^P_{t-m}$	$m^P_{\cdot}$	$m^P_{t-1}$	$m^P_{t+1}$	$m^P_{T-m}$	$m^P_{T+1}$
$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$
$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$
$t+m+2$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$
$t+m+3$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$
$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$
$T+2$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$
$T+3$	$s^P_{t-2h+1}$	$\cdot$	$s^P_{t-h}$	$s^P_{t-h+1}$	$\cdot$	$s^P_{t-m}$	$s^P_{\cdot}$	$s^P_{t-1}$	$s^P_{t+1}$	$s^P_{T-m}$

<sup>a</sup> Le processus de révision prévoit  $m$  révisions successives de type courant ( $2^P, \dots, m^P, r$ ), une révision aux jalons due aux contrôles effectués toutes les  $h$  périodes ( $b^P$ ) et une révision extraordinaire à l'occasion de laquelle on reconstruit une série rétrospective à partir du temps  $T+2$  (Source: Biggri 1984, p.24).

la différence entre une estimation provisoire et l'estimation correspondante révisée est une pure mesure de la différence dans l'erreur entre les deux estimations successives. Toutefois, quand le processus de révision se base sur des informations de plus en plus complètes et sur des méthodes d'évaluation de plus en plus affinées, comme il arrive justement dans les pays développés, il est raisonnable de considérer l'estimation finale comme la plus proche de la vraie valeur inconnue, et dans des buts de comparaison, de la traiter comme valeur de référence. La connaissance des caractéristiques et du comportement des discordances entre estimations provisoires et révisées est d'une utilité certaine. De fortes et fréquentes différences constituent, en effet, un symptôme préoccupant quant à la qualité des données de base et/ou des méthodes d'estimation employées dans les évaluations naturellement pas que de petites discordances soient nécessaires. (Ceci ne signifie naturellement pas que de petites discordances soient nécessaires, mais cela garantit la qualité des données. Le fait qu'une évaluation préliminaire ne soit pas révisée peut simplement vouloir dire que nous n'avons pas d'éléments suffisants pour l'améliorer, abstraction faite de sa fiabilité.) Si en plus les différences présentent des traits de systématicité, le fait peut représenter, en même temps, un utile avertissement tant pour l'utilisateur, qui éventuellement adoptera des correctifs, que pour le producteur, qui pourra en déduire des suggestions pour améliorer les estimations provisoires.

Un compte-rendu des nombreuses analyses statistiques des discordances entre estimations provisoires et révisées qui se sont succédées, est donné par Trivellato (1986a). D'autres récentes contributions sont de Zarnowitz (1982), Mankiw, Runkle et Shapiro (1984), Mankiw et Shapiro (1986), McNees (1986), Mork (1987) et Lefrançois (1988). Sur ce thème la littérature est donc considérable. Mais certaines questions, tant méthodologiques que substantielles, concernant la manière de mener les analyses, ont été partiellement négligées ou résolues de façon quelquefois peu satisfaisante. En particulier, ceci vaut pour: (a) l'identification d'un dessin des analyses de la précision des estimations provisoires, qui soit cohérent avec les caractéristiques du processus de révision; (b) la définition des mesures élémentaires appropriées des discordances entre estimations provisoires et révisées; le choix d'indices synthétiques convenables, qui soient capables de nous informer de la précision globale d'une série d'estimations provisoires, et, si le processus de révision se déroule en plusieurs étapes, qui puissent être décomposés de façon à pouvoir vérifier la convergence des estimations provisoires vers la valeur finale; (c) l'examen des implications des erreurs dans les estimations provisoires sur les séries dérivées, notamment sur les agrégats à prix constants et les déflateurs implicites. Les trois sections suivantes sont consacrées, dans l'ordre, à ces problèmes. La section 5, enfin, rapporte de façon sommaire les résultats de quelques analyses empiriques sur les discordances entre des estimations provisoires et révisées, menées comparativement sur des données de comptabilité nationale du Canada, des Etats-Unis et de l'Italie.

## 2. UN SCHEMA THEORIQUE DU PROCESSUS DE REVISION

Un schéma suffisamment général du processus de révision des agrégats économiques est représenté dans le tableau 1, qui permet une visualisation immédiate du lien entre la période de référence et la période de publication des différentes estimations. De l'examen du schéma émergent trois types d'estimations et de révisions: (a) Celles qui sont placées sur la diagonale principale et sur les diagonales inférieures, qui décrivent un processus de révision courant se déroulant en  $m$  pas, allant de l'estimation préliminaire  $1p$  à l'estimation définitive  $1r$  ("définitive" par rapport au processus de révision courant). (b) Celles placées horizontalement ( $br$ ), qui incorporent les ajustements aux jalons, où la reconstruction de la série se fait en référence à une période qui va d'un étalonnage au suivant.

# L'évaluation des erreurs dans les données de comptabilité nationale: les estimations provisoires et révisées

LUIGI BIGGERI et UGO TRIVELLATO<sup>1</sup>

## RÉSUMÉ

Cet article présente et discute de façon critique les récents développements dans l'évaluation de la fiabilité des estimations provisoires de comptabilité nationale. On introduit d'abord un schéma théorique du processus de révision des estimations d'un agrégat, et on considère les implications qu'il a sur le dessin des analyses des erreurs dans les données provisoires. Une attention particulière est ensuite portée au choix des mesures élémentaires des erreurs et d'indices synthétiques de précision convenables, et à l'impact des révisions sur les agrégats à prix constants et les déflateurs implicites. Enfin, les résultats de quelques analyses empiriques sur les discordances entre des estimations provisoires et révisées, menées comparativement sur des données de comptabilité nationale du Canada, des États-Unis et de l'Italie, sont synthétiquement présentés.

MOTS CLÉS: Révision des données; comptes nationaux; indices de précision; agrégats à prix constants.

## 1. INTRODUCTION

L'évaluation de la fiabilité des estimations de comptabilité nationale est importante à cause de l'influence de ces estimations sur les analyses et les décisions de politique économique. Elle est en même temps difficile à effectuer, en raison des très nombreuses et complexes sources d'erreur dont elles sont affectées.

Il est donc problématique d'arriver à un unique critère de fiabilité, à la fois convaincant au niveau conceptuel et praticable. Il apparaît en revanche raisonnable et possible d'établir de nombreux critères partiels, convenables pour mesurer les aspects principaux de la fiabilité. Pour une revue des différents critères nous renvoyons le lecteur à Novak (1975) et Trivellato (1987). L'objectif de cet article est de présenter et de discuter de façon critique certains développements récents dans l'analyse des erreurs dans les estimations provisoires de comptabilité nationale.

Pour les agrégats de comptabilité nationale, ainsi que pour beaucoup d'autres agrégats et indicateurs économiques, des estimations provisoires sont initialement publiées, et sont ensuite soumises à différentes révisions.

Le processus de révision est déterminé avant tout par l'exigence d'assurer des informations rapides, et, en même temps, par le temps requis pour le recueil et l'élaboration de l'ensemble des données utilisées couramment pour l'estimation des agrégats (''actualité contre précision'', selon l'efficace polarisation de Wilton et Smith 1974). Typiquement ceci donne lieu à un processus de révision caractérisé par une estimation préliminaire et par une suite de révisions de routine à périodicité rapprochée. De temps à autre il peut ensuite arriver que des modifications soient apportées aux classifications et schémas de comptabilité et/ou dans les enquêtes statistiques de base et dans les méthodes d'estimation, pour améliorer la précision ou la pertinence des données. Ceci ouvre la voie à de nouvelles évaluations, et nous parlons dans ce cas de révisions occasionnelles ou extraordinaires.

La particularité des erreurs dans les estimations provisoires réside dans le fait qu'elles sont présentes dans ces estimations et, par définition, éliminées dans les estimations révisées. Elles peuvent donc être mesurées en comparant les deux estimations. Certes, à strictement parler,

<sup>1</sup> Luigi Biggieri, Dipartimento Statistico, Università di Firenze, Italia; Ugo Trivellato, Dipartimento di Scienze Statistiche, Università di Padova, Italia.





où  $n = \sum n_h$  et  $g = \sum g_{hj}$  (nota: si  $Xq = (1, \dots, 1)'$  pour un vecteur quelconque  $q$  de dimension  $K$ , alors  $g = 0$ ).  
 Lorsqu'il manque des variables explicatives, les éléments diagonaux de  $eqm'$  risquent d'être affectés d'un biais par excès. Le raisonnement que nous venons d'exposer est conforme à celui appliqué par Wolter (1985) pour les estimateurs de la variance pour strates regroupées selon la théorie des sondages à base de plans.

### BIBLIOGRAPHIE

JOHNSTON, J. (1972). *Econometric Methods*, (deuxième édition). New York: McGraw Hill.  
 KOTT, P.S. (1991). A model-based look at linear regression with survey data. *American Statistician*, à paraître.  
 KOTT, P.S. (1990a). What does performing a linear regression on survey data mean? *Proceedings of the Section on Survey Research Methods, American Statistical Association*, à paraître.  
 KOTT, P.S. (1990b). The design consistent regression estimator and its conditional variance. *Journal of Statistical Planning and Inference*, 24, 287-296.  
 SHAH, B.V., HOLT, M.M., et FOLSOM, R.E. (1977). Inference about regression models from sample survey data. *Bulletin of the International Statistical Institute*, 47, 43-57.  
 WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.  
 ZELLNER, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association*, 57, 348-368.

additionnelle qui ait un lien avec la variable explicative présumée manquante de l'autre équation (ce qui n'est pas étonnant étant donné que lorsqu'on a introduit un terme  $x_{7i}$  dans le membre de droite de l'équation (12), le coefficient estimé correspondant était inférieur à la valeur estimée de la racine carrée de l'erreur quadratique moyenne).

### 7. SYSTÈME SIMULTANÉ

Dans un système simultané, on retrouve quelques-unes des colonnes de la matrice des variables dépendantes,  $Y$ , (voir équ. (3)) dans le membre de droite de la  $g$ -ième équation (voir (5)). Formellement, nous pouvons écrire

$$y = Y^{(\cdot)}\alpha + X\beta + u + v \quad \text{ou} \quad y = Z\delta + u + v,$$

où

$$Y^{(\cdot)} = \begin{bmatrix} Y^{(1)} \\ Y^{(2)} \\ \vdots \\ Y^{(g)} \end{bmatrix},$$

$$Z = (Y^{(\cdot)}, X), \quad \text{et} \quad \delta = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}.$$

La plupart des colonnes de  $Y^{(g)}$  sont des vecteurs nuls. Les autres (pas plus que  $G-1$  colonnes) proviennent de la matrice  $Y$  de l'équation (3). Définissons  $Y^{(g)}$  comme  $X(X'WX)^{-1}X'WY^{(g)}$ . Remplaçons ensuite  $X$  dans l'équation (5) par  $Z = (Y^{(\cdot)}, X)$  et refaisons les opérations précédentes. L'équation (7) donne  $\delta^{W.MCO}$ , ce qui tient des doubles moindres carrés, tandis que l'équation (8) donne  $\delta^{W.MCG}$ , qui tient des triples moindres carrés. En ce qui concerne l'estimation de l'erreur quadratique moyenne, on applique le raisonnement qui a permis d'obtenir l'équation (9).

### 8. ANALYSE

Par cet article, nous visions à montrer comment les méthodes décrites dans les ouvrages portant sur la théorie des sondages à base de plans – en particulier, la régression pondérée (pour échantillon) et l'estimateur de variance "de linéarisation" – pouvaient être appliquées à l'estimation d'un système d'équations linéaires. Notre analyse nous a amenés à faire une constatation quelque peu surprenante: lorsqu'on estime les paramètres d'un système assujéti à des conditions, il se peut que les équivalents pondérés de l'estimateur par les MCO et de l'estimateur par les MCG ne servent pas à estimer la même chose. À bien y penser toutefois, cette constatation n'est pas si étonnante qu'on le croit. S'il manque des variables explicatives dans notre modèle de travail, peut-être nous ne connaissons pas toujours assez le vrai modèle pour pouvoir appliquer des conditions aux paramètres.

Il importe de préciser que l'équation (9) peut servir à estimer l'erreur quadratique moyenne d'estimateurs de paramètres même lorsqu'il ne manque aucune variable explicative. L'avantage de cette formule par rapport aux formules classiques est qu'elle est conçue en fonction de l'hétéroscédasticité et de corrélations complexes entre les données simples (mais à l'intérieur des u.p.é.). Néanmoins, lorsqu'il ne manque aucune variable explicative, l'estimateur ci-dessous présente tous les avantages de l'estimateur de l'équation (9) et est généralement plus efficace:



Tableau 1

Estimateurs pour les paramètres de l'équation (11)

MCO	$y_{1i} = 0.268x_{1i} - 0.92x_{2i} + u_{1i} + v_{1i}$	(.044) (3.95)
Pondéré	$y_{1i} = 0.191x_{1i} + 12.15x_{2i} + u_{1i} + v_{1i}$	(.075) (9.95)
MCG - pondéré	$y_{1i} = 0.197x_{1i} + 10.26x_{2i} + u_{1i} + v_{1i}$	(0.71) (6.97)

Les chiffres entre parenthèses représentent la racine carrée de l'erreur quadratique moyenne.

non nécessaire),  $T^2$  est distribué asymptotiquement selon la loi chi carré avec 2 degrés de liberté. Nous ne pouvons rejeter l'hypothèse nulle à un seuil de 0.1. Cependant, comme la valeur de  $T^2$  est beaucoup plus élevée que 2, il est fortement permis de croire qu'il manque une variable explicative. Par conséquent, il conviendrait d'utiliser l'estimateur par régression pondéré plutôt que l'estimateur par les MCO.

Le tableau 1 donne les estimations des coefficients de l'équation (11) calculées à l'aide de l'estimateur par les MCO et de son équivalent pondéré. Bien que l'écart entre l'estimateur pondéré de  $\beta_2$  et la valeur nulle ne soit pas statistiquement significatif à un seuil de 0.1, nous conservons cet estimateur dans le modèle car sa valeur est supérieure à la valeur estimée de la racine carrée de l'erreur quadratique moyenne correspondante. Cela rejoint notre argument en faveur de l'estimateur par régression pondéré.

Il faut toutefois noter la perte d'efficacité qui découle de l'utilisation d'un estimateur pondéré. En effet, la valeur estimée de la racine carrée de l'erreur quadratique moyenne de l'estimateur de  $\beta_2$  fait plus que doubler lorsqu'on passe de l'estimateur par les MCO à son équivalent pondéré (nota: dans les deux cas, la racine carrée de l'erreur quadratique moyenne a été calculée au moyen de l'équation (9)).

Il est possible d'accroître l'efficacité de l'estimateur pondéré en introduisant une seconde équation pour l'exploitation agricole et en la jumelant à l'équation (11) pour obtenir un système à estimer. Soit

$$y_{2i} = x_{1i}\beta_3 + u_{2i} + v_{2i}, \tag{12}$$

où  $y_{2i}$  est le rapport de la superficie ensemencée au maïs à la superficie de terre labourable pour l'exploitation  $i$  si cette exploitation a des terres labourables; dans le cas contraire, ce rapport est nul.

L'existence d'un système d'équations ne change rien aux estimations calculées au moyen de l'estimateur par les MCO pondéré ni à la valeur estimée de la racine carrée de l'erreur quadratique moyenne correspondante (tableau 1). Cependant, nous pouvons aussi calculer désormais la valeur estimée de  $\beta_1$  et  $\beta_2$  au moyen de l'estimateur par les MCG pondéré; les valeurs pertinentes figurent aussi dans le tableau 1. Il convient de noter, à cet égard, la réduction d'environ 30% de la valeur estimée de la racine carrée de l'erreur quadratique moyenne pour les MCO et de l'estimateur par les MCG servant à estimer la même chose.

La valeur  $T^2$  pour un test permettant de comparer les versions pondérées de l'estimateur par les MCO et de l'estimateur par les MCG pour le vecteur  $(\beta_1, \beta_2)$  est 0.97. Comme cette valeur est sensiblement inférieure à 2, il est permis de croire que les deux estimateurs donnent le même résultat. Autrement dit, l'une des équations ne renferme aucune variable explicative

Un critère général permettant de vérifier si

$$\hat{\beta}^{(1)} = \sum_{j=1}^f \sum_{h=1}^h \{C^{(1)} * D_{hj} y^*\} \quad \text{et} \quad \hat{\beta}^{(2)} = \sum_{j=1}^f \sum_{h=1}^h \{C^{(2)} * D_{hj} y^*\}$$

sont égaux est

$$T^2 = [\hat{\beta}^{(1)} - \hat{\beta}^{(2)}]' A^{-1} [\hat{\beta}^{(1)} - \hat{\beta}^{(2)}], \quad (10)$$

où

$$A = \sum_{h=1}^H \frac{1}{n_h} \left[ \sum_{j=1}^f d_{hj} d_{hj}' - \frac{1}{n_h} \left( \sum_{j=1}^f d_{hj} \right) \left( \sum_{j=1}^f d_{hj}' \right) \right],$$

$$d_{hj} = C^{(1)} * D_{hj} r_{hj(1)} - C^{(2)} * D_{hj} r_{hj(2)}, \text{ et } r_{hj(f)} = y_{hj} - X_{hj} \hat{\beta}^{(f)}.$$

Suivant l'hypothèse nulle, le critère  $T^2$  est une variable aléatoire distribuée asymptotiquement selon une loi de chi carré avec  $K$  degrés de liberté. Compte tenu de l'importance que nous accordons à la robustesse, il semble prudent de s'interroger sur la validité de l'hypothèse nulle lorsque  $\text{prob}(\chi^2_{(K)} > T^2)$  est largement inférieure au niveau normal de 0.1 ou de 0.05 mais non lorsque  $T^2$  est inférieur à son espérance,  $K$ .

## 6. EXEMPLE

Considérons l'exemple ci-dessous, construit à l'aide de données de l'enquête agricole de juin 1989 du National Agricultural Statistics Service. La série de données pertinente, analysée antérieurement dans Kott (1990a), est décrite brièvement ci-après. On a tout d'abord prélevé 17 unités primaires d'échantillonnage à même 4 strates. Ces unités ont ensuite fait l'objet d'un sous-échantillonnage qui a permis d'obtenir un échantillon global de 252 exploitations agricoles. Bien qu'il s'agisse d'un échantillon aléatoire, la probabilité de sélection n'était pas la même pour toutes les exploitations. Supposons que nous voulons estimer les paramètres  $\beta_1$  et  $\beta_2$  de l'équation suivante:

$$y_{1i} = x_{1i} \beta_1 + x_{2i} \beta_2 + u_{1i} + v_{1i}, \quad (11)$$

où  $i$  désigne une exploitation agricole;  $y_{1i}$  est le rapport de la superficie ensemencée au soja à la superficie de terre labourable pour l'exploitation  $i$  si cette exploitation a des terres labourables; dans le cas contraire, ce rapport est nul;  $x_{1i}$  est égal à 1 si l'exploitation  $i$  a des terres labourables et égal à 0 dans le cas contraire;  $x_{2i}$  est le quotient de la superficie de terre labourable de l'exploitation  $i$  par 10,000.

(Nota: le fait d'éliminer de l'équation de régression toutes les exploitations échantillonnées qui n'ont pas de terre labourable n'influe aucunement sur l'estimation des paramètres mais peut avoir des effets sur l'estimation de l'erreur quadratique moyenne.)

Si nous posons  $\hat{\beta}^{(1)}$  dans l'équation (10) comme l'estimateur par les MCO non pondéré pour le vecteur  $(\beta_1, \beta_2)'$  et  $\hat{\beta}^{(2)}$  comme l'estimateur pondéré, nous obtenons une valeur  $T^2$  de 4.58. Suivant l'hypothèse nulle que l'estimateur par les MCO et son équivalent pondéré permettent d'arriver au même résultat (ce pour quoi  $u_{1i} \equiv 0$  est une condition suffisante mais

est un estimateur convergent selon le plan de  $b_{MCG}$ . Comme  $b_{MCO}$  et  $b_{MCG}$  (et pour les mêmes raisons),  $\hat{\beta}_{W \cdot MCO}$  et  $\hat{\beta}_{W \cdot MCG}$  sont égaux lorsque le système d'équations n'est assujéti à aucune condition.

Si  $\hat{\beta}_{W \cdot MCO}$  et  $\hat{\beta}_{W \cdot MCG}$  sont convergents selon le plan, ils sont aussi des estimateurs quasi-rapport au modèle de l'équation (6)).

(Dans cet article, la propriété d'être sans biais est toujours définie par  $\lim_{M \rightarrow \infty} X'U/M = 0_K$ ). (Dans cet article, la propriété d'être sans biais est toujours définie par

4. ESTIMATION DE L'ERREUR QUADRATIQUE MOYENNE

Supposons que nous avons un plan d'échantillonnage qui prévoit  $H$  strates,  $n_h$  u.p.é. tirées de la strate  $h$ , et  $m_{hj}$  données simples tirées de l'u.p.é.  $h_j$ . Les estimateurs  $\hat{\beta}_{W \cdot MCO}$  et  $\hat{\beta}_{W \cdot MCG}$  sont tous deux de la forme  $\hat{\beta} = Cy$ . Sans perte de généralité, nous pouvons réexprimer ces estimateurs sous la forme  $\hat{\beta} = Cy^*$ , où  $y^* = (y_{11}', \dots, y_{Hn_H}')$  renferme uniquement les éléments qui correspondent aux données simples échantillonnées, et  $y_{hj}$  est le vecteur des valeurs- $y \cdot G \times m_{hj}$  relatives aux données simples de l'u.p.é.  $h_j$ . Définissons  $r^*$  et  $r_{hj}$  de la même manière que  $y^*$  et  $y_{hj}$ .

Soit  $D_{hj}$  une matrice diagonale de zéros et de uns telle que  $D_{hj}y^* = (0', \dots, y_{hj}', \dots, 0')$  et posons  $g_{hj} = C'D_{hj}r^*$ . Par une extension directe de l'estimateur "de linéarisation", fondé sur un plan, nous pouvons exprimer l'estimateur de l'erreur quadratique moyenne de  $\hat{\beta} = Cy^*$  sous la forme:

(9) 
$$eqm = \sum_H \frac{n_h}{n_h - 1} \left[ \sum_{j=1}^{n_h} g_{hj} g_{hj}' - \frac{1}{n_h} \left( \sum_{j=1}^{n_h} g_{hj} \right) \left( \sum_{j=1}^{n_h} g_{hj}' \right) \right].$$

Suivant des contraintes modérées pour le plan d'échantillonnage, eqm est quasi-non biaisé lorsque  $U$  (d'après l'équation (3))  $\equiv 0_{M \times G}$  et  $V$  répond à la propriété suivante:

$$\left\{ \begin{array}{l} 0 \text{ lorsque } s \text{ et } t \text{ ne viennent pas de la même u.p.é.} \\ > 0 \text{ dans le cas contraire.} \end{array} \right. | E(v_{sg} v_{tf})$$

Voir Kott (1991) pour la démonstration dans le cas où  $G = 1$ ; la démonstration pour  $G > 1$  est triviale. L'erreur quadratique moyenne de l'estimateur demeure à un niveau raisonnable lorsque  $U \neq 0_{M \times G}$  (voir Kott 1990a).

5. CRITÈRES STATISTIQUES

Soit  $\hat{\beta}_{r \cdot MCO}$  et  $\hat{\beta}_{r \cdot MCG}$  les équivalents non pondérés de  $\hat{\beta}_{W \cdot MCO}$  et  $\hat{\beta}_{W \cdot MCG}$ , que l'on obtient en remplaçant  $W$  par  $S$  dans les équations (7) et (8). On cherche souvent à savoir si l'utilisation de poids d'échantillonnage a un effet quelconque. Cela nous amène à vérifier si il existe une différence significative entre  $\hat{\beta}_{r \cdot MCO}$  et  $\hat{\beta}_{W \cdot MCO}$ , autrement dit si les deux estimateurs servent à estimer la même chose.

Une fois que l'importance de l'utilisation de poids a été établie, on veut aussi savoir s'il existe une différence significative entre  $\hat{\beta}_{W \cdot MCO}$  et  $\hat{\beta}_{W \cdot MCG}$ , autrement dit si la relation  $\lim_{M \rightarrow \infty} X'U/M = 0_{K \times G}$  se vérifie de telle sorte que les deux estimateurs en question servent à estimer la même chose.



$$\hat{\sigma}_{gf} = r_{\cdot g, W, f} / \sum_{i=1}^M w_i, \text{ et } r = y - X\hat{\beta}_{W, MCO},$$

$$\begin{aligned} \hat{\beta}_{W, MCO} &= (X' [I_G \otimes W] [\hat{\Sigma}^{-1} \otimes I_M] X)^{-1} X' [I_G \otimes W] y \\ &= (X' [X^{-1} X] W [\hat{\Sigma}^{-1} \otimes I_M] X)^{-1} X' [X^{-1} X] W y, \end{aligned} \quad (8)$$

$0_K$ . Dans les mêmes conditions, l'estimateur par les MCO pondérée est un estimateur convergent selon le plan de  $b_{MCO}$ , ce qui implique que  $\lim_{m \rightarrow \infty} (\hat{\beta}_{W, MCO} - b_{MCO}) =$

$$\hat{\beta}_{W, MCO} = (X' [I_G \otimes W] X)^{-1} X' [I_G \otimes W] y \quad (7)$$

(voir Kott 1990b, 1991), l'estimateur par les MCO pondérée Il est facile de montrer que pour de nombreuses populations et de nombreux plans de sondage

lorsque  $p_i = m/M$  pour tous les  $i$ ,  $W = S$ . matrice des poids d'échantillonnage, où  $m$  représente la taille de l'échantillon. Notons que tillon et 0 dans le cas contraire. Enfin, posons  $W = (m/M) P^{-1} S = \text{diag}\{w_i\}$  comme la donnée simple  $i$ . Soit  $S = \text{diag}\{s_i\}$ , où  $s_i = 1$  si la donnée simple  $i$  est incluse dans l'échantillon aléatoire de la population. Soit  $P = \text{diag}\{p_i\}$ , où  $p_i$  est la probabilité de sélection de la Supposons maintenant que nous n'observons des valeurs de variables que pour un échan-

### 3. ESTIMATION À L'AIDE DE DONNÉES D'ENQUÊTE

expression dont la valeur tend vers zéro lorsque  $M$  augmente, suivant l'hypothèse plus robuste mais pas nécessairement suivant l'hypothèse moins robuste.

$$\sum_{\cdot g} X_{\cdot g, f}' \left( \sum_{\cdot g} \sigma_{fg}^2 u_{\cdot f} \right) / M = \sum_{\cdot g} \sum_{\cdot f} \sigma_{fg}^2 X_{\cdot g, f}' u_{\cdot f} / M,$$

$$E(b_{MCO} - \beta) \propto X' [\hat{\Sigma}^{-1} \otimes I_M] u / M =$$

lignes  $\{(g - 1)M + 1\}$  à  $\{gM\}$  de  $X$  et  $\hat{\Sigma}^{-1} = \{\sigma_{fg}^2\}$ ; alors, Pour comprendre le phénomène, définissons  $X_{\cdot g}$  comme la matrice  $M \times K$  formée des

(6) représente une équation unique et non un système d'équations.  $\lim_{m \rightarrow \infty} X' u / M = 0_K$ , ce qui est plus en rapport avec le modèle élargi de Kott (1991) lorsque Seulement,  $b_{MCO}$  peut ne pas être quasi-non biaisé suivant l'hypothèse moins rigoureuse que l'équation  $\lim_{m \rightarrow \infty} X' U / M = 0_{K \times G}$  se vérifie, comme nous l'avons supposé à l'origine. si  $u \neq 0_{MG}$ ,  $b_{MCO}$  et  $b_{MCO}$  sont tous deux des estimateurs quasi-non biaisés de  $\beta$  lorsque assujettie à aucune condition (voir de nouveau Johnston 1972, p. 240). Dans le cas contraire, Il est notoire que  $b_{MCO}$  et  $b_{MCO}$  sont égaux lorsque la matrice des paramètres en (3) n'est

de la formule  $\hat{\sigma}_{gf} = r_{\cdot g, f} / M$ , où  $r_{\cdot g} = y_{\cdot g} - X_{\cdot g}' b_{MCO}$ .

que, les éléments de  $\hat{\Sigma}$  doivent être estimés à partir de l'échantillon, par exemple au moyen  $I_M$  est la matrice unité  $M \times M$ , est le meilleur estimateur linéaire sans biais. Dans la pratique, est un estimateur sans biais de  $\beta$  mais  $b_{MCO} = (X' [\hat{\Sigma}^{-1} \otimes I_M] X)^{-1} X' [\hat{\Sigma}^{-1} \otimes I_M] y$ , où

Lorsque  $u \equiv 0_{MG}$  et  $\text{Var}(v) = \hat{\Sigma} \otimes I_M$  (où  $\hat{\Sigma} = \{\sigma_{st}\}$ ), alors  $b_{MCO} = (X' X)^{-1} X' y$  s'agira plus alors d'une matrice diagonale par blocs).

on peut extraire le second du vecteur  $\beta$  de l'équation (6) et corriger  $X$  en conséquence (il ne colonne correspondante de la matrice  $X$ . Si deux éléments d'une même ligne de  $\beta$  sont égaux, est nul, on peut supprimer cet élément du vecteur  $\beta$  de l'équation (6) en même temps que la Toutefois, lorsque c'est l'inverse,  $K < GK$ . Par exemple, si on sait qu'un des éléments de  $\beta$

où  $Y$  est la matrice  $M \times G$  des variables dépendantes observées (la  $i$ -ième ligne de  $Y$  renferme les variables dépendantes qui se rapportent à la  $i$ -ième donnée simple),

$X$  est la matrice  $M \times K$  des variables indépendantes (ou explicatives) observées (la  $i$ -ième ligne de  $X$  renferme les variables indépendantes qui se rapportent à la  $i$ -ième donnée simple),

$\beta$  est la matrice  $K \times G$  des paramètres,

$U$  est une matrice  $M \times G$  qui satisfait l'équation  $\lim_{M \rightarrow \infty} X'U/M = 0_{K \times G}$  (un  $G$ -vecteur de valeurs nulles) – cela suppose l'existence d'un processus générateur de données simples qui pourrait, en principe, générer de telles données à l'infini (voir Kott 1991),

$V$  est une matrice  $M \times G$  de variables aléatoires telle que  $E(V) = 0_{M \times G}$  (une matrice de valeurs nulles) et  $E(V_{is}V_{jt}) = \sigma_{st}(i)$ .

Nous savons tous que si  $U \equiv 0_{M \times G}$ ,  $E(V_{is}V_{jt}) = 0$  pour  $i \neq j$ , et  $\sigma_{st}(i) = \sigma_{st}$  pour tous  $i$ , alors

$$(4) \qquad B_{MCO} = (X'X)^{-1}X'Y$$

est le meilleur estimateur linéaire sans biais de  $\beta$  (voir, par exemple, Johnston 1972, p. 240). Cela signifie que la  $g$ -ième colonne de  $B_{MCO}$  (appelons-la  $B_{\cdot g}$ ) est le meilleur estimateur linéaire sans biais de  $\beta_{\cdot g}$ , où

$$(5) \qquad y_{\cdot g} = X\beta_{\cdot g} + u_{\cdot g} + v_{\cdot g},$$

et  $y_{\cdot g}$ ,  $u_{\cdot g}$  et  $v_{\cdot g}$  sont les  $g$ -ième colonne de  $Y$ ,  $U$  et  $V$  respectivement. L'équation (5) peut être considérée comme la  $g$ -ième équation du système représenté par l'équation (3).

Désignons la matrice  $U$  de l'équation (3) comme la matrice des *variables explicatives présumées manquantes*. Dans les analyses de régression classiques (c.-à-d. fondées sur un modèle), on suppose généralement que la portion des variables dépendantes qui ne peut être décrite par une combinaison linéaire des variables indépendantes est purement aléatoire. Toutefois, dans la présente analyse, nous reprenons le raisonnement de Kott (1991) et reconnaissons la possibilité de variables explicatives manquantes non aléatoires. Notons que même lorsque  $U \neq 0_{M \times G}$ ,  $B_{MCO}$  est quasi (c.-à-d. asymptotiquement) sans biais.

## 2.2 Système (vraisemblablement) assujéti à des conditions

Il est plus difficile de réaliser une estimation efficace lorsque des contraintes s'appliquent à certains éléments de  $\beta$ , par exemple lorsqu'on sait que  $\beta_{hg}$  est nul ou que  $\beta_{hj}$  égale  $\beta_{hj}$ . Dans cet article, nous nous intéressons à un système d'équations (vraisemblablement) assujéti à des conditions qui peut être représenté par l'équation suivante:

$$(6) \qquad y = X\beta + u + v,$$

où  $y = (y_{\cdot 1}', y_{\cdot 2}', \dots, y_{\cdot G}')'$ ,  $u$  et  $v$  sont définis de la même manière,  $X$  est une matrice  $MG \times K$ ,  $\beta$  est un vecteur  $K \times 1$  et  $K \leq GK$ . Par définition,  $\lim_{M \rightarrow \infty} X'u/M = 0_K$ . Lorsque la matrice  $\beta$  de l'équation (3) n'est pas assujéti à des contraintes,  $K = GK$  et

$$X = \begin{bmatrix} X & & \\ & X & \\ & & \ddots \\ & & & X \end{bmatrix}$$

Pour notre second exemple, supposons que nous avons un échantillon d'entreprises qui fabriquent un produit donné,  $y$ , à partir de deux intrants,  $x_1$  et  $x_2$ , dont le prix unitaire est  $p_1$  et  $p_2$  respectivement. Les économistes vont souvent supposer que les entreprises se situent toutes au même niveau technologique (à quelque différence près). Étant donné  $p_1$ ,  $p_2$  et  $y$ , chaque entreprise choisira  $x_1$  et  $x_2$  de manière à minimiser le coût total,  $c = p_1x_1 + p_2x_2$ . Supposons que nous puissions exprimer la relation entre  $p_1$ ,  $p_2$  et  $y$  et le coût total  $c$  au moyen de l'équation suivante (en moyenne):

(1) 
$$\log(c) = b_0 + b_1\log(p_1) + b_2\log(p_2) + b_3\log(y).$$

La théorie économique nous dit que devant une équation de coût implicite comme l'équation (1), une entreprise avisée déterminera le niveau de  $x_1$  de manière que

(2) 
$$x_1p_1/c = b_1.$$

Naturellement, pour que nous puissions estimer les équations (1) et (2), nous devons prévoir une structure stochastique. Pour des raisons de simplicité, nous supposons que les deux équations décrivent fidèlement le comportement de toutes les entreprises, moyennant des erreurs aléatoires indépendantes (d'une entreprise à l'autre) et identiquement distribuées. Outre la forte possibilité d'une corrélation entre les termes d'erreur des deux équations pour une entreprise en particulier, notons qu'il y a un coefficient ( $b_1$ ) commun aux deux équations. Lorsque nous sommes en présence d'un système d'équations linéaires dont les coefficients sont assujettis à des conditions, il est peu utile de recourir aux méthodes de régression linéaire fondées sur un plan dont fait état Kott (1990a). C'est pourquoi les méthodes exposées dans cet article seront envisagées uniquement dans l'optique de l'approche fondée sur un modèle de Kott (1991), même si bon nombre d'entre elles s'inspirent d'approches fondées sur un plan.

Dans la section 2, nous exposons le modèle théorique pour l'estimation d'un système d'équations linéaires à l'aide de données d'une population. Dans la section suivante, nous présentons l'équivalent pondéré des estimateurs par les MCO et les MCG d'une population pour un système d'équations linéaires. La section 4 traite de l'estimation robuste de l'erreur quadratique moyenne de ces versions pondérées par une simple généralisation de l'estimateur de variance "de linéarisation" (voir, par exemple, Shah, Holt et Folsom 1977). Dans la section 5, nous examinons une méthode générale permettant de construire des critères statistiques qui peuvent servir à déterminer, entre autres choses, si les versions pondérées des estimateurs par les MCO et les MCG estiment toutes deux le même paramètre. Nous présentons un exemple simple dans la section 6 tandis que dans la section suivante, nous entreprenons d'étendre la méthodologie exposée précédemment à ce que les économètres appellent les "systèmes simultanés". En ce qui a trait à la version stochastique de l'équation 1 par exemple, de nombreux économistes croient que l'on devrait considérer  $\log(y)$  comme une variable aléatoire et que l'on peut supposer  $\log(c)$  fixe. Ces conditions ont pour effet de créer un biais de simultanéité si l'on ne recourt pas à des méthodes comme les doubles ou les triples moindres carrés (voir Johnston 1972, p. 341-420). Enfin, la section 8 renferme une brève analyse.

## 2. ESTIMATION POUR UNE POPULATION

### 2.1 Système non assujéti à des conditions

Supposons que nous avons une population avec  $M$  données simples. Chacune de ces données (i) se rapporte à  $G + K$  variables observées qui satisfont le modèle suivant:

(3) 
$$Y = X\beta + U + V,$$



# Estimation d'un système d'équations linéaires à l'aide de données d'enquête

PHILLIP S. KOTT<sup>1</sup>

## RÉSUMÉ

Dans cet article, nous élaborons un modèle permettant d'estimer un système d'équations linéaires à l'aide de données d'enquête. La théorie classique des sondages, axée sur des plans, nous est peu utile ici, bien que quelques-unes des techniques élaborées en vertu de cette théorie puissent être incorporées à des méthodes d'estimation robuste fondées sur un modèle. Les estimateurs de variance qui ont la forme de l'estimateur de "linéarisation" à équation unique sont quasi-non biaisés suivant de nombreuses structures d'erreur complexes. En outre, l'absence possible de variables explicatives peut être compensée par l'introduction de poids d'échantillonnage dans l'estimation par régression. Dans certains cas toutefois, l'absence de variables explicatives peut rendre incertaine l'estimation d'un système d'équations.

**MOTS CLÉS:** Poids d'échantillonnage; variable explicative présumée manquante; robuste; quasi-non biaisé.

## 1. INTRODUCTION

Kott (1991) a montré que les techniques à base de plans conçues pour estimer les équations de régression linéaire simples pouvaient servir dans des analyses fondées sur un modèle. Il a montré notamment que l'utilisation de la régression pondérée (pour échantillon) pouvait compenser l'absence possible de variables explicatives et que l'estimateur de variance dit "de linéarisation" pouvait produire des estimateurs quasi-non biaisés de l'erreur quadratique moyenne pour de nombreuses structures de variance complexes.

Dans le présent article, nous étendons les résultats de cette analyse à l'estimation d'un système ou "groupe" d'équations linéaires, sujet de très grand intérêt pour les économètres (voir, par exemple, Johnston (1972), p. 238-241). Deux exemples élémentaires seront peut-être éclairants pour le lecteur qui n'est pas habitué aux méthodes économétriques ou à leur équivalent.

Supposons que nous avons un échantillon d'exploitants agricoles et que nous voulons déterminer la relation entre la superficie ensemencée au soja et la taille de l'exploitation agricole – dont le terme d'erreur était corrélé avec celui de la relation originale. L'estimation globale d'un tel groupe d'équations a été désignée par Zellner comme la "régression sans relation apparente". Étrangement, pour que l'estimateur par les moindres carrés généralisés (MCG) que propose Zellner donne des résultats différents de ceux de l'estimateur par les MCO, certaines équations doivent contenir des variables explicatives que l'on ne retrouve pas dans d'autres équations. On peut aussi imaginer que toutes les équations renferment les mêmes variables explicatives mais que dans certains cas, des coefficients sont nuls.

<sup>1</sup> Phillip S. Kott, Special Assistant for Economic Survey Methods, U.S. Bureau of the Census, Room 3061-3, Washington, DC, 20233, E.-U.

## ANNEXE

Dans la présente annexe, nous utilisons la notation de la sous-section 2.3.2.

Nous déterminons d'abord la formule pour  $U_d(t)$ , la taille du domaine des unités "disparues" utilisée dans l'EERH pour la  $t$ -ième sélection.

Rappelons-nous de l'expression  $U_d(t) = w(t)[\hat{v}(t) + \hat{n}_d(t)]$ . Au moment de la  $t$ -ième mise à jour, on trouve  $d(t+1)$  unités "disparues" dans l'échantillon. Nous pouvons donc remplacer  $\hat{n}_d(t)$ , le nombre estimé d'unités disparues présentes dans l'échantillon, par  $d(t+1)$ , afin d'obtenir une mise à jour d' $U_d(t)$ , c'est-à-dire  $N_d(t+1)$ :

$$(1.1) \quad N_d(t+1) = w(t)[\hat{v}(t) + d(t+1)].$$

La formule (1.1) utilise la correction en fonction des disparitions du mois précédent. La valeur initiale de  $\hat{v}$  est  $\hat{v}(0) = 0$  (voir la remarque après la formule (2.9) et la définition d' $\hat{n}$ ) et  $w(0) = f^{-1}$ . La taille du domaine des unités "disparues" pour la  $(t+1)$ -ième sélection mensuelle est désormais estimée par:

$$(1.2) \quad U_d(t+1) = \max(N_d(t+1) - D(t+1) - d(t+1), 0).$$

Notons que  $U_d(t+1)$  peut être calculée à partir de (2.5) lorsque  $U_d(t+1)$  est connue, et réciproquement. On obtient une autre forme pour  $U_d(t+1)$  par des calculs récursifs de la façon suivante. Supposons que  $U_d(t)$  est connue avant la  $t+1$ -ième sélection du  $t$ -ième échantillon (rappelons-nous que l'expression  $t=0$  est utilisée pour désigner la première sélection). Alors,  $U_d(t)$  est aussi connue et peut être utilisée pour calculer  $P_d(t)$ , la probabilité de prélever une unité disparue parmi les unités hors échantillon (voir la formule (2.6)). Cette probabilité est ensuite utilisée pour calculer le nombre voulu d'unités qui devraient être introduites dans l'échantillon par renouvellement comme il est décrit à la sous-section 2.3.2, ainsi que le nombre prévu d'unités actives présentes dans l'échantillon au moment de la  $(t+1)$ -ième sélection,  $\hat{n}_t(t)$ .

Le poids utilisé dans l'estimation pour la sélection suivante est alors déterminé par  $U_d(t)/\hat{n}_t$ . Après la sélection, la  $(t+1)$ -ième mise à jour a lieu et on constate que le nombre réel d'unités actives présentes dans l'échantillon est  $n_t(t+1)$ . On peut estimer la taille du domaine des unités actives pour la prochaine sélection à l'aide de la formule suivante:

$$U_d(t+1) = \min \left\{ U_d(t) \frac{\hat{n}_t(t)}{n_t(t)} + 1, N(t+1) + B(t+1) + n_t(t+1) \right\}$$

et ainsi de suite. Pour amorcer le processus, notons que le poids utilisé dans la première estimation est déterminé par l'équation  $w(0) = f^{-1}$  et que, après la première mise à jour,

$$U_d(1) = \min \{ w(0) \times n_t(1) + B(1), N(1) \}.$$

## BIBLIOGRAPHIE

- COTTREL-BOYD, T.M., DUNN, M.R., HUNTER, G.E., et SRINATH, K.P. (1980). Development of the redesign of the Canadian establishment based employment surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 8-15.
- SCHIOPU-KRATINA, I., et SRINATH, K.P. (1986). The methodology of the Survey of Employment, Payroll and Hours. Working Paper No. BSM-D-86-010E, Statistique Canada.
- STATISTIQUE CANADA (1970). *Classification des activités économiques*. N° 12-501 au catalogue, Ottawa: Statistique Canada.

Dans l'EBRH, la valeur estimée des unités "actives"  $U_i$  est définie par:

$$(3.7) \quad U_i = wn_i,$$

le poids étant déterminé à l'aide de la formule (3.1).

La définition du poids dans l'EBRH (voir la fin de la section 2.3) suppose que:

$$(3.8) \quad U_i = \begin{cases} U_i^* & \text{si } n_i \leq fN \\ N & \text{si } n_i > fN. \end{cases}$$

L'estimation du nombre total d'employés dans l'EBRH est définie par:

$$(3.9) \quad Y_i = U_i^* Y_i^*.$$

Un estimation du nombre total d'employés fondée sur (3.5) est obtenue par la formule:

$$(3.10) \quad Y_i^* = U_i^* Y_i^*.$$

L'estimateur fondé sur la formule (3.8) est biaisé et, par conséquent, l'estimateur du nombre total d'employés dans l'EBRH l'est aussi.

Toutefois, l'erreur quadratique moyenne de l'estimateur fondé sur (3.9), conditionnellement à la valeur de  $n_i$ , est inférieure à celle de l'estimateur fondé sur (3.10), conditionnellement à la valeur de  $n_i$ . Nous allons maintenant tenter de prouver cette affirmation.

On s'aperçoit facilement que, pour chaque résultat particulier, le biais  $B_i$  de l'estimateur fondé sur (3.9) et conditionnellement de la valeur de  $n_i$  est déterminé par:

$$(3.11) \quad B_i = (U_i - N_i) Y_i^*.$$

De même, nous obtenons pour l'estimateur fondé sur (3.10) l'équation suivante:

$$(3.12) \quad B_i^* = (U_i^* - N_i) Y_i^*.$$

Nous montrons que l'erreur quadratique moyenne conditionnelle de l'estimateur fondé sur (3.9) est inférieure à celle de l'estimateur fondé sur (3.10). Le même résultat est aussi valide pour les erreurs quadratiques moyennes inconditionnelles. Nous prenons comme condition la taille d'échantillon réelle des unités "actives"  $n_i$ . À partir des formules (3.8) à (3.10),  $\text{Var}[U_i^* Y_i^* | n_i] - \text{Var}[U_i Y_i | n_i] = [n_i^2 f^{-2} - N^2] 1\{n_i > fN\} \text{Var}[Y_i | n_i]$ . Notons que  $n_i^2 f^{-2} - N^2 > 0$  dans l'ensemble  $\{n_i > fN\}$ . Nous comparons maintenant  $[B_i^*]^2$  et  $[B_i]^2$ :

$$[B_i^*]^2 - [B_i]^2 = 1\{n_i f^{-1} > N\} \{ [U_i^*]^2 Y_i^2 - [N]^2 Y_i^2 - (N - N_i)^2 Y_i^2 \}.$$

Cependant,  $U_i^* - N_i = f^{-1} n_i - N_i > N - N_i$  si  $n_i f^{-1} > N$ .

Par conséquent,  $[B_i^*]^2 - [B_i]^2 \geq 0$ . Étant donné que  $\text{E}[\text{EQM}[U_i^* Y_i^* | n_i] = \text{Var}[U_i^* Y_i^* | n_i] + [B_i^*]^2$  et que  $\text{E}[\text{EQM}[U_i Y_i | n_i] = \text{Var}[U_i Y_i | n_i] + [B_i]^2$ , la comparaison terme par terme nous amène à conclure que:

$$\text{MSE}[U_i^* Y_i^* | n_i] \geq \text{MSE}[U_i Y_i | n_i].$$

Cette importante propriété motive le choix de l'estimateur fondé sur (3.9) plutôt que de l'estimateur fondé sur (3.10) pour évaluer le nombre total d'employés dans l'EBRH.



des disparitions (voir (2.11)), on calcule le poids initial qui est attribué à chaque unité présente dans l'échantillon aux fins de l'estimation à l'aide de la formule (voir (2.12)):

$$(3.1) \quad w = \frac{n}{N} + \phi.$$

Toutefois, si  $t$  unités à valeurs aberrantes sont présentes dans l'échantillon ( $t \geq 1$ ), on modifie alors ce poids en accordant à chaque unité à valeur aberrante un poids de 1 et en attribuant aux unités restantes dans la partie à tirage partiel de l'échantillon le poids calculé suivant la formule

$$(3.2) \quad w' = \frac{n}{N - t} + \phi - t.$$

Supposons que  $S$  représente l'ensemble des unités présentes dans l'échantillon. Si  $Y(n)$  désigne la valeur du nombre d'employés correspondant à l'unité  $n$  dans l'échantillon, on détermine alors l'estimateur du nombre total d'employés dans la strate par la formule

$$(3.3) \quad Y = \sum_{n \in S} w(n)Y(n),$$

où  $w(n) = 1$  si  $n$  est une unité à valeur aberrante ou une unité à tirage complet et où  $w(n) = w'$  dans le cas de toutes les autres unités présentes dans l'échantillon (voir (3.2)). Des estimations des totaux à n'importe quel niveau d'aggrégation supérieur à celui de la strate sont obtenues en additionnant les estimations des totaux pour toutes les strates dans le niveau d'aggrégation en question.

Nous allons analyser certaines des propriétés de l'estimateur défini par le dernier terme du côté droit de la formule (3.3). Étant donné que, aux fins de l'estimation, les unités à valeurs aberrantes peuvent être considérées comme des unités à tirage complet, nous pouvons remplacer, pour simplifier,  $w'$  par  $w$ , ce dernier étant défini par la formule (3.1). Il faut alors modifier  $N$  et  $n$  en conséquence (voir (3.2)).

Soient  $N_t$  la taille du domaine des unités "actives" qu'on retrouve dans la population de la strate et  $n_t$  le nombre d'unités "actives" présentes dans l'échantillon. Supposons aussi que  $Y_t$  représente le nombre d'employés moyen pour les unités "actives" présentes dans l'échantillon. Étant donné que seules les unités "actives" contribuent à l'estimation du nombre total d'employés, on obtient une estimation du total pour la cellule grâce à la formule:

$$(3.4) \quad Y_t = U_t Y_t.$$

Nous considérons le cas  $fN > m$ . Dans la formule (3.4),  $U_t$  est une valeur estimée de  $N_t$ , le nombre d'unités actives dans la population, et est définie par:

$$(3.5) \quad U_t = \frac{f}{1} n_t.$$

Dans la formule (3.5),  $f$  désigne la fraction de sondage de la strate qui est gardée fixe. L'estimateur fondé sur les valeurs de  $U_t$  est non biaisé, c'est-à-dire:

$$(3.6) \quad E(U_t) = N_t.$$

Étant donné que l'estimateur fondé sur (3.5) est non biaisé,  $U_t$  peut excéder  $N$  dans certains cas, puisque  $U_t$  peut avoir des valeurs peu élevées pour compenser.

Il faut noter qu'une sous-représentation des créations peut se produire dans certains cas. Par exemple, lorsque, dans une strate quelconque, aucune unité ne doit être introduite dans l'échantillon par renouvellement, alors, pour le mois en question, les créations ne seront pas représentées dans l'échantillon prélevé de cette strate. Toutefois, cette situation survient d'habitude lorsque la taille de la population est peu élevée et que, à la longue, on peut s'attendre à ce que, en moyenne, les créations soient représentées correctement.

Les  $b$  créations réellement prélevées sont attribuées au hasard aux groupes de renouvellement 1 à 12 dans l'échantillon, ce qui se traduit, en moyenne, à  $b/12$  de créations qui sont attribuées à chacun de ces groupes. Ainsi, on s'assure que la probabilité de voir une création être supprimée de l'échantillon par renouvellement est la même que pour  $n$  importe quelle autre unité.

Afin de garder constante la répartition par âge des unités des groupes i) à iii) (voir la section 2.1), les créations non échantillonnées seront attribuées au hasard au groupe des unités AS et à celui des unités NAS. Cela signifie que, si  $N'$  est le nombre d'unités NAS, dans le groupe des unités AS et  $N''$  le nombre d'unités NAS, créations non échantillonnées seront alors attribuées au groupe des unités AS et  $N''(B - b)/(N' + N'') + N''(B - b)/(N' + N'')$  créations non échantillonnées au groupe des NAS.

### 3. ESTIMATION

#### 3.0 Introduction

Dans le présent chapitre, nous décrivons la méthode d'estimation du "nombre total d'employés rémunérés".

Comme nous l'avons indiqué à la section 2.0, seule la fiabilité des estimations du nombre total d'employés au niveau d'aggrégation des groupes d'activité économiques (échelle provinciale) est définie à l'avance. Les estimations relatives à des caractéristiques autres que le nombre total d'employés présentent différents degrés de fiabilité. Par exemple, on s'attend à ce que les estimations de la rémunération hebdomadaire moyenne soient plus fiables que celles du nombre total d'employés.

Le niveau d'aggrégation le plus bas auquel les estimations sont publiées est le niveau de la CAF et de la province, mais les unités de base pour la production des estimations sont les strates. Une valeur aberrante dans l'ERH est une valeur observée dans la partie à tirage partiel de l'échantillon, qui est supérieure à une valeur préalable.

Le poids de chaque strate est d'abord calculé comme dans la section 2.3, puis il est corrigé en fonction des valeurs aberrantes dans la strate en question. Des estimations du nombre total d'employés sont établies pour chaque strate à l'aide des poids corrigés. Il n'y a pas de correction en fonction de la non-réponse, puisque les valeurs correspondant aux non-répondants font l'objet d'une imputation. On obtient l'estimation du nombre total d'employés pour chaque strate en additionnant les totaux suivants:

- i) le nombre total d'employés dans les unités à tirage complet;
  - ii) le nombre total d'employés dans les unités à valeurs aberrantes;
  - iii) la somme des valeurs pondérées du nombre d'employés dans les unités à tirage partiel, à l'exclusion des unités à valeurs aberrantes.
- Étant donné que le poids qui est attribué à chaque unité à valeurs aberrantes est de un, les unités à valeurs aberrantes sont traitées comme des unités à tirage complet aux fins de l'estimation et ne sont donc pas utilisées dans l'estimation de la variance.

#### 3.1 Estimation du total pour une caractéristique

Considérons une strate particulière pour un mois donné dans l'enquête. Soient  $N$  la taille de la partie à tirage partiel de la population de la strate et  $n$  le nombre d'unités à tirage partiel présentes dans l'échantillon pour le mois en question. Si  $v$  correspond à la correction en fonction

## 2.4 Échantillonnage des créations

Comme nous l'avons mentionné précédemment, de nouvelles unités sont enregistrées dans la base chaque mois. Étant donné qu'on estime que ces nouvelles unités (créations) peuvent différer des "anciennes", nous avons conçu une stratégie particulière pour représenter adéquatement les créations dans l'échantillon.

L'idéal serait que, si  $B$  créations sont enregistrées durant le mois courant et si  $f$  désigne la fraction de sondage de la strate,  $b = fB$  créations devraient être présentes dans l'échantillon durant le mois en question. Les créations échantillonnées sont attribuées au hasard aux groupes de renouvellement décrits dans la section 2.2. De cette façon, on s'assure que la probabilité de suppression de l'échantillon par renouvellement est la même pour les nouvelles unités et pour les "anciennes", de sorte que la répartition par âge des unités à l'intérieur de l'échantillon est la même que celle des unités hors échantillon.

En utilisant la notation de la section précédente, soient  $n_i$  le nombre d'unités qui doivent être introduites dans l'échantillon par renouvellement au moment de la sélection mensuelle de l'échantillon, à l'exclusion des créations, et  $N'$  le nombre d'unités dans le groupe des unités AS (groupe ii) de la section 2.2). La stratégie relative aux créations consiste en une sélection en deux phases. On prélève des créations ou des unités "plus anciennes" pour former échantillon préliminaire, duquel un échantillon est ensuite tiré. Cette façon de procéder est nécessaire parce qu'un groupe, normalement celui des créations, est petit. Ensuite,  $n'$  unités sont pré-échantillonnées du groupe des unités AS et placées dans le groupe des créations (voir (2.15)). Si:

$$(2.13) \qquad \frac{N'}{n_i} > f,$$

un certain nombre de créations,  $b'$ , sont pré-échantillonnées du groupe des créations et placées dans le groupe des unités AS.

La valeur de  $b'$  est déterminée par la formule:

$$(2.14) \qquad b' = \frac{b}{N'} n_i.$$

L'inégalité dans (2.13) garantit que  $b' \leq B$ , ce qui signifie que le groupe des créations est suffisamment grand et que le pré-échantillonnage peut avoir lieu. À partir de l'échantillon préliminaire de  $N' + b'$  unités,  $n_i + b'$  unités sont ensuite échantillonnées.

Le choix de  $b'$ , tel qu'il est déterminé par la formule (2.14), garantit que le nombre prévu de créations présentes dans l'échantillon est celui qui est souhaité, puisque la probabilité de prélever une création de l'échantillon préliminaire est  $b' / (b' + N')$ ; par conséquent, le nombre prévu de créations lorsque  $n_i + b'$  unités sont tirées sans remise de cet échantillon préliminaire est  $(n_i + b') b' / (b' + N')$ , qui, d'après (2.14), est égal à  $b$ .

Dans la situation complémentaire, lorsque la formule (2.13) n'est pas valide, nous avons que:

$$(2.15) \qquad n' = n_i / f \leq N'.$$

Alors,  $n'$  unités "plus anciennes" sont prélevées du groupe des unités AS et placées dans le groupe des créations. Notons que, dans cette situation,  $b'$ , tel que défini par la formule (2.14), est supérieur à  $B$  et que, par conséquent, on ne peut pas appliquer la première méthode.

Nous calculons maintenant le nombre prévu d'unités "anciennes" dans l'échantillon. Étant donné que la probabilité de prélever une unité "ancienne" est désormais  $n' / (n' + B)$  et que  $n_i + b'$  unités sont tirées de l'échantillon préliminaire de  $n' + B$  unités et placées dans l'échantillon, le nombre prévu d'unités "anciennes" est  $(n_i + b') n' / (n' + B)$ , qui est égal à  $(n_i + b)$ .



où  $\hat{u}(t)$  sera déterminé ultérieurement (voir (2.10)). La valeur de  $\hat{u}(t)$  représente les "disparitions non réelles" qui ont été ajoutées à l'échantillon pour représenter correctement les unités "disparues" dans la population. Notons que, lorsque le premier échantillon est prélevé ou qu'on procède à un nouveau prélèvement, un tel redressement n'est pas nécessaire, c'est-à-dire que  $\hat{u}(0) = 0$ .

Afin de trouver une formule pour  $\hat{u}(t)$ , nous utilisons la formule (2.5) dans le numérateur de (2.9) et la formule (2.8) dans le dénominateur. D'après (2.7) et (2.8), nous devons aussi avoir:

$$\frac{\hat{U}_d^a(t)}{1} = \frac{\hat{u}_d(t) + \hat{u}(t)}{f}.$$

Etant donné que  $\hat{u}_d(t)$  est égal à  $P_d(t)n_i(t)$ , nous obtenons, à partir de l'équation ci-dessus:

$$\hat{U}_d^a(t) = \frac{f}{1} [\hat{u}_d(t) + \hat{u}(t)] \quad \text{ou} \quad \hat{u}(t) = f \hat{U}_d^a(t) - \hat{u}_d(t). \quad (2.10)$$

On peut faire la correction en fonction des disparitions à l'aide de la formule:

$$\hat{v}(t) = \begin{cases} 0 & \text{si } \hat{u}(t) > -\hat{u}_d(t) \\ \hat{u}(t) & \text{si } \hat{u}(t) \geq -\hat{u}_d(t) \end{cases} \quad (2.11)$$

et on calcule le poids utilisé dans l'estimation suivant la formule

$$w(t) = \frac{n(t) + \hat{v}(t)}{N(t)}. \quad (2.12)$$

Notons que le poids dans (2.12) est défini à l'aide d'un estimateur et qu'il s'agit donc d'une

variable aléatoire.

L'utilisation du poids défini par la formule (2.12) suppose que l'estimateur du nombre

d'unités actives dans la population, laquelle est établie par l'équation  $\hat{U}_d^a(t) = w(t)\hat{u}_d(t)$ , ne dépasse pas  $N(t)$ , la taille de la population au moment du  $t$ -ième échantillonnage.

Définissons:

$$\hat{U}_d^a(t) = w(t) [\hat{v}(t) + \hat{u}_d(t)].$$

D'après (2.10) et (2.11), il s'ensuit que  $\hat{U}_d^a(t) \geq 0$  et que sa valeur minimale est 0 lorsque  $\hat{v}(t) - \hat{u}_d(t) = 0$ . D'après (2.5), la valeur maximale de  $\hat{U}_d^a(t)$  est alors  $N(t)$ .

On traitera dans la section 3.1 du fait que la restriction selon laquelle l'estimateur des unités actives soit tronqué à  $N(t)$  à des répercussions sur l'estimation. L'estimateur  $\hat{U}_d^a(t)$  est calculé récursivement (voir l'annexe) à l'aide des formules (2.10) et (2.11), et  $\hat{v}(0) = \hat{u}(0) = 0$ .

Il faut noter que la formule (2.11) est légèrement différente de celle qui permet de faire la

correction en fonction des disparitions dans l'ERH. Premièrement, pour simplifier, nous

n'avons pas considéré ici les cas où l'on doit utiliser la taille minimale de l'échantillon,  $m$ . Dans

de tels cas, il n'est pas approprié de se servir de la fraction de sondage  $f$  dans (2.10). Dans

l'ERH, on utilise dans tous les cas le poids du mois précédent (2.10) au lieu de  $f$ . Deuxième-

ment, la correction en fonction des disparitions dans l'ERH est toujours considérée positive. La formule (2.11) montre qu'elle peut être négative, tant qu'elle est supérieure à  $-\hat{u}_d(t)$ . Il

est très rare que cette situation se présente ou, de façon plus générale, lorsque  $\hat{u} \leq 0$ .

L'annexe présente une formule qui permet de calculer l'estimateur des unités actives et qui ne nécessite pas la correction en fonction des disparitions.

unités actives doivent être tirées du groupe des unités admissibles à la sélection et être ajoutées à l'échantillon avec les créations prélevées. Compte tenu de l'existence d'un nombre inconnu d'unités inactives dans la population et de l'arrondissement au nombre entier, il faut que:

$$n_1(t) = \min \left( \left\lfloor \frac{\ell(t)}{1 - P_d(t)} + 0.5 \right\rfloor, n''(t) \right),$$

plus d'unités soient introduites dans le même échantillon par renouvellement,  $P_d(t)$  étant définie par la formule (2.6) et  $n''(t)$  étant égale à  $N(t) - n_o(t)$ .

En calculant  $n_1(t)$ , nous avons supposé qu'il n'y a pas d'unités inactives parmi les créations, de sorte que le facteur d'élargissement  $[1 - P_d(t)]^{-1}$  ne s'applique qu'aux unités "plus anciennes" dans le groupe des unités AS.

On détermine la taille de l'échantillon,  $n(t)$ , pour le  $t$ -ième mois à l'aide de la formule: (2.8)

$$n(t) = \max\{n_1(t) - n_o(t) + n_1(t) + b(t), m\},$$

un laps de temps considérable. Des  $n_1(t)$  unités qui sont introduites dans l'échantillon par renouvellement, on s'attend à trouver des unités inactives (dont le nombre est calculé à l'aide de la formule  $\hat{n}_d(t) = P_d(t)n_1(t)$ ) et des unités actives (dont le nombre est déterminé par la formule  $\hat{n}_a(t) = n_1(t) - \hat{n}_d(t)$ ). Ainsi, des  $n(t)$  unités présentes dans l'échantillon après les  $t$ -ièmes prélèvements et renouvellement mensuels, on prévoit qu'un certain nombre d'entre elles (calculé en utilisant la formule  $\hat{n}_a(t) = n_1(t) - n_o(t) + \hat{n}_a(t) + b(t)$ ) seront "actives", et elles représentent les  $U_i(t)$  unités du domaine des unités "actives" au taux approprié  $f$  (voir (2.7) et (2.8)), lorsque  $n(t) > m$  dans (2.8).

### 2.3.3 Détermination des poids

Le poids  $w(t)$  utilisé aux fins de l'estimation pour le  $t$ -ième mois est exprimé en fonction de la taille de la population et de l'échantillon au passage  $t$ . Toutefois, l'utilisation de  $N(t)/n(t)$  comme poids aux fins de l'estimation pourrait amener une surestimation des unités actives dans la population. En effet,  $n(t)$  dans la formule (2.8) a été choisi pour que le nombre prévu d'unités actives présentes dans l'échantillon égale le nombre cherché. Il se peut que le nombre d'unités disparues présentes dans l'échantillon prélevé comme il est décrit ci-dessus ne représente pas la taille du domaine des unités disparues au taux approprié. Dans la formule (2.8),  $n_1$  est tiré de la population en général et on s'attend alors à ce qu'il preserve la proportion entre le domaine des unités "disparues" et celui des unités "actives". Aucune disparition ne devrait être trouvée parmi les créations et, par conséquent,  $b(t)$  représente bien le sous-groupe des créations. Il reste toutefois dans l'échantillon  $n_1(t) - n_o(t)$  unités d'une sélection précédente, après le renouvellement et les mises à jour de l'échantillon. Il est probable que la proportion des disparitions parmi ces unités soit de beaucoup inférieure à la proportion correspondante dans la population en général, malgré le fait que les mises à jour sont fondées sur des renseignements provenant de sources autres que l'enquête. La valeur de  $N(t)/n(t)$  doit alors être corrigée en fonction de la sous-représentation des unités "disparues" dans l'échantillon. On obtient ainsi, lorsque  $n(t) > m$ ,

$$\frac{N(t)}{1} = \frac{n(t) + \hat{n}(t)}{f}, \quad (2.9)$$

Ici,  $[a]$  représente le nombre entier le plus grand qui n'est pas supérieur à  $a$ . La constante 0,5 est utilisée pour obtenir une meilleure approximation au moment de l'arrondissement. Supposons que  $d(t)$  unités sont éliminées de l'échantillon (disparitions à l'intérieur de l'échantillon) et  $D(t)$  unités du reste de la population de la strate (disparitions hors échantillon). Supposons aussi que  $B(t)$  nouvelles unités (des créations) sont enregistrées durant le même intervalle.

Par conséquent, la taille de la population de la cellule au moment de la  $t$ -ième sélection est calculée à l'aide de la formule:

$$(2.3) \quad N(t) = N(t-1) - d(t) - D(t) + B(t).$$

Puisque les mises à jour ne sont pas exhaustives, on s'attend à ce que la population compte des unités inactives (disparues) non décelées.

Soit  $n_t(t)$  le nombre d'unités actives laissées dans l'échantillon du mois précédent (après les mises à jour), c'est-à-dire:

$$(2.4) \quad n_t(t) = n(t-1) - d(t).$$

Nous supposons qu'il n'y a pas d'unités disparues non décelées dans l'échantillon à ce point-ci. Nous pouvons croire que la population de la strate est répartie dans deux domaines: le domaine des unités "actives" et celui des unités "disparues". La taille du domaine des unités "disparues" n'est pas connue, mais on peut calculer un estimateur  $U_d(t)$  d'après les renseignements fournis par l'échantillon et les mises à jour (voir l'annexe). Soit  $U_d(t)$  un estimateur du nombre d'unités "disparues" non décelées dans la population au moment de la  $t$ -ième sélection mensuelle. Alors:

$$(2.5) \quad N(t) = U_t(t) + U_d(t),$$

où  $U_t(t)$  est l'estimateur du nombre d'unités actives. La probabilité de choisir une unité disparue lorsqu'on prélève une unité au hasard parmi les unités hors échantillon est définie à l'aide de la formule:

$$(2.6) \quad P_d(t) = \min \left\{ \frac{U_d(t)}{N(t) - n_t(t)}, 1 \right\}.$$

On détermine le nombre voulu d'unités actives dans l'échantillon par la formule:

$$(2.7) \quad n_t^*(t) = f U_t(t).$$

On calcule la taille de l'échantillon de remplacement de façon à s'assurer que le nombre prévu d'unités actives dans l'échantillon après la sélection est  $n_t^*(t)$ .

Supposons maintenant qu'au moment des  $t$ -ièmes sélection et renouvellement des  $n_t(t)$  unités actives dans l'échantillon,  $n_o(t)$  unités peuvent être supprimées de l'échantillon par renouvellement.

Etant donné qu'il reste  $n_t(t) - n_o(t)$  unités actives dans l'échantillon, on a besoin de  $n_t^*(t) - n_t(t) + n_o(t)$  plus d'unités actives dans l'échantillon pour le  $t$ -ième mois.

Afin de représenter adéquatement les créations dans l'échantillon,  $b(t) = fB(t)$  créations doivent être prélevées au hasard et incluses dans l'échantillon.

Par conséquent:

$$l(t) = \max(n_t^*(t) - n_t(t) + n_o(t) + b(t), 0),$$



Les unités qui sont laissées dans l'échantillon sont attribuées à un groupe d'attente, lequel est divisé en sous-groupes. Un sous-groupe est composé d'unités qui ont toutes été supprimées de l'échantillon le même mois. Le groupe d'attente contient 12 sous-groupes dans chaque strate. Le temps que chaque unité passe en-dehors de l'échantillon est ainsi consigné afin de s'assurer que les unités ne seront pas prélevées de nouveau avant au moins 12 mois. Les unités qui ont passé la période minimale requise dans le groupe des unités "non admissibles à la sélection" sont transférées dans le groupe des unités "admissibles à la sélection", et on leur attribue ainsi une probabilité positive de ré-échantillonnage.

Pour résumer, toute la population à tirage partiel comporte à  $n$  importe quel moment donné quatre groupes d'unités. Il s'agit:

- i) des unités qui sont présentes dans l'échantillon pour le mois en question;
- ii) des unités admissibles à la sélection (AS);

- iii) des unités dans le groupe d'attente qui ont été supprimées de l'échantillon par renouvellement moins de 12 mois auparavant et qui ne sont pas admissibles à la sélection (NAS); et
- iv) des créations, c'est-à-dire des unités qui n'ont pas été enregistrées dans la base auparavant.

Le processus mensuel de sélection et de renouvellement comporte un échange d'unités entre ces groupes. Certaines unités quittent le groupe i) pour se retrouver dans le groupe iii) et de nouvelles unités entrent dans le groupe i) en provenance du groupe ii), après que certaines créations échantillonnées du groupe iv) ont été transférées au groupe ii) et que le reste des créations ont été attribuées aux groupes ii) et iii) après sélection. Cette façon de procéder permet de s'assurer que l'échantillon est représentatif de la population à  $n$  importe quel mois donné.

## 2.3 Détermination de la taille de l'échantillon et des poids

### 2.3.1 Mises à jour mensuelles

La base de sondage contient un grand nombre d'unités qui sont inactives ou hors du champ d'observation, qui ne sont plus en exploitation, etc. À part le fardeau de conserver un nombre croissant d'unités inactives dans la base, il est probable que les estimateurs fondés sur des échantillons tirés d'une telle population aient une variance élevée, à cause du fait que l'échantillon renferme une proportion élevée de valeurs observées nulles. L'idéal serait d'éliminer toutes ces unités de la base de sondage avant que l'échantillon mensuel ne soit prélevé. La base est mise à jour chaque mois, entre les opérations mensuelles de sélection et de renouvellement. C'est pour cette raison que les indices que nous utilisons pour désigner les créations et les disparitions dans la base sont d'une unité plus élevés que ceux employés pour désigner la taille de l'échantillon au moment du tirage précédant la mise à jour. Par exemple, après le prélèvement du premier échantillon, supposons qu'il y a  $n(0)$  unités dans l'échantillon, desquelles on constate par la suite que  $d(1)$  unités sont des unités disparues. Alors  $D(1)$  désigne le nombre d'unités disparues dans la partie hors échantillon de la population et  $B(1)$  le nombre d'unités enregistrées en tant que créations le mois en question. En calculant la taille d'échantillon voulue pour le mois suivant,  $n(1)$ , on doit tenir compte de ces mises à jour (voir (2.3)) ainsi que de la taille de la population au moment de la sélection du premier échantillon,  $N(0)$ .

### 2.3.2 Détermination de la taille de l'échantillon

La population d'une strate à tirage partiel donnée est une fonction du temps et sera désignée, disons, par  $N(t)$ , où  $t$  est un nombre entier positif qui s'accroît d'une unité d'un mois à l'autre. On détermine la taille d'échantillon voulue pour chaque mois par la formule:

$$n'(t) = [fN(t) + 0.5].$$

On définit la fraction de sondage dans chaque strate par la formule:

(2.1) 
$$f = \max \left( f', \frac{1}{100} \right),$$

où  $f'$  est calculée au niveau de la superstrate correspondante. Afin de réduire l'instabilité des estimations qui est causée par les valeurs peu élevées de la fraction de sondage, on a décidé de fixer à 1/100 la fraction de sondage minimale pour toutes les strates.

Les calculs détaillés de la taille de l'échantillon au niveau des strates sont donnés à la section 2.3 (voir la façon dont est déterminée la formule (2.8)). Un échantillon est prélevé systématiquement de chaque strate.

## 2.2 Le plan de renouvellement

Le renouvellement de l'échantillon (remplacement périodique partiel de l'échantillon) dans l'EBRH sert principalement à réduire le fardeau de réponse. D'après des enquêtes antérieures, il est apparu que le taux de réponse moyen dans les strates où il n'y avait pas de renouvellement était de beaucoup inférieur à celui que l'on constatait dans les strates où il y avait renouvellement. De plus, la présence dans l'échantillon d'un grand nombre d'unités pendant deux mois consécutifs rehausse la fiabilité des estimations de la variation d'un mois à l'autre par rapport aux estimations de la variation sur des échantillons mensuels indépendants. Le renouvellement de l'échantillon dans chaque strate doit être effectué suivant certaines contraintes comme, par exemple, laisser les unités hors de l'échantillon pendant une certaine période après qu'elles ont été supprimées de celui-ci par renouvellement.

L'échantillon mensuel est composé de 14 groupes numérotés de 0 à 13. Le groupe 0 contient les unités à tirage complet de la strate. Les groupes 1 à 13 sont appelés "groupes de renouvellement". Les numéros 1 à 12 attribués aux groupes de renouvellement indiquent les mois où les unités autres que les créations ont été introduites dans l'échantillon par renouvellement. Par exemple, le groupe de renouvellement 1 renferme surtout des unités qui ont été échantillonnées en janvier et des créations, le groupe de renouvellement 2 contient principalement des unités qui ont été introduites dans l'échantillon en février et des créations, et ainsi de suite. Le groupe de renouvellement 13 se compose d'unités qui ont été présentes dans l'échantillon pendant 12 mois. Ces unités sont les plus anciennes en ce qui a trait au temps passé dans l'échantillon et peuvent être supprimées par renouvellement. Chaque mois, des créations sont échantillonnées et attribuées au hasard aux groupes de renouvellement.

Au moment de la sélection et du renouvellement mensuels, toutes les unités dans le mois de référence sont transférées dans le groupe de renouvellement 13. En février, par exemple, toutes les unités du groupe de renouvellement 2 sont transférées dans le groupe de renouvellement 13. Un groupe de remplacement est choisi à partir d'unités "admissibles à la sélection" et d'unités nouvellement enregistrées (des créations). Les unités du groupe de remplacement (à l'exception de 1/12 des créations) sont ensuite placées dans le groupe de renouvellement 2, et elles ne pourront être supprimées de l'échantillon par renouvellement avant au moins 12 mois. Si l'on dispose de suffisamment d'unités pour un groupe de remplacement, les unités contenues dans le groupe 13 sont éliminées de l'échantillon et ne peuvent pas être ré-échantillonnées avant au moins 12 mois. Sinon, certaines unités du groupe 13 sont conservées dans l'échantillon jusqu'à ce que l'on ne dispose pas d'un nombre suffisant d'unités en-dehors de l'échantillon pour former un groupe de remplacement. Cette mesure sert à maintenir la taille de l'échantillon à un minimum ou à s'assurer de disposer d'un échantillon suffisamment grand pour établir des estimations qui répondent à des normes de fiabilité préétablies. De cette façon, en général au moins 11/12 des unités demeurent dans la partie à tirage partiel de l'échantillon pendant deux mois consécutifs.



1.1 Définitions préliminaires

Voici la définition de certains des termes utilisés dans le présent article.

- i) Etablissement – Plus petite unité étant une entité juridique capable de déclarer tous les éléments des données de base sur les activités économiques. L'établissement est l'unité statistique pour l'EERH. Nous allons utiliser le terme unité pour établissement.
- ii) Unité déclarante (pouvant fournir des renseignements sur l'emploi) – Aux fins de la compilation de statistiques géographiques détaillées, l'établissement est souvent subdivisé en unités déclarantes en fonction principalement de l'emplacement et, parfois, d'autres aspects comme, entre autres, la liste de paye.
- iii) Classification des activités économiques (CAE) – 1970. On attribue à chaque établissement un code CAE selon la nature de son activité. Ces codes sont définis dans le Manuel de la CAE (document de référence [5]).

Dans l'EERH, le nombre total d'employés rémunérés lié à une unité est la caractéristique choisie pour mesurer la taille de ladite unité.

L'EERH comporte quatre groupes de taille dont les limites sont définies de la façon suivante: 0 à 19 employés (groupe de taille 1), 20 à 49 employés (groupe de taille 2), 50 à 199 employés (groupe de taille 3) et 200 employés et plus (groupe de taille 4).

iv) On définit la superstrate en fonction d'une division d'activités économiques, d'une province et d'un groupe de taille. Puisqu'il y a 16 divisions d'activités économiques, 12 provinces et quatre groupes de taille, il y a donc 768 superstrates.

v) On définit la strate en fonction d'un code CAE à trois chiffres, d'une province et d'un groupe de taille. Il s'agit du niveau le plus détaillé pour lequel on établit des estimations.

vi) La partie à tirage complet de la population se compose d'unités qui sont toutes incluses dans l'échantillon avec certitude. Elle contient des unités du groupe de taille 4 et des unités de la population définies à l'avance. La partie à tirage partiel de la population comprend les unités restantes qui sont soumises à l'échantillonnage, comme il est décrit aux sections suivantes.

2. SÉLECTION ET RENOUVELLEMENT DE L'ÉCHANTILLON

2.0 Détermination de la taille de l'échantillon et méthode de répartition

Dans l'EERH, on détermine la taille de l'échantillon à tirage partiel pour chaque groupe d'activités économiques au niveau d'une province selon un coefficient désigné de variation de l'estimation du nombre total d'employés pour ce groupe d'activités économiques et pour cette province. La taille d'échantillon cherchée et les fractions de sondage sont calculées au niveau de la superstrate à l'aide d'une répartition proportionnelle à la taille, cette dernière étant le nombre total d'employés. Les fractions de sondage sont gardées constantes d'un mois à l'autre. On trouve plus de détails sur la méthode de répartition dans le document de référence [3]. La sélection réelle est faite au niveau de la strate. En raison du minimum d'unités que doit compter l'échantillon à ce niveau, le nombre d'unités échantillonnées est plus élevé que la taille d'échantillon voulue pour chaque groupe d'activités économiques au niveau provincial (voir (2.2)).

2.1 Sélection de l'échantillon

Considérons maintenant une strate particulière. Soient  $N$  la taille de la partie à tirage partiel de la population et  $n$  la taille de l'échantillon à tirage partiel dans cette strate. Tandis que la répartition de l'échantillon entre les superstrates est proportionnelle à la taille, la répartition au niveau des strates dans chaque superstrate est essentiellement proportionnelle au nombre total d'unités retrouvées dans la partie à tirage partiel de cette strate.



# Renouvellement de l'échantillon et estimation dans l'Enquête sur l'emploi, la rémunération et les heures de travail

IOANA SCHIOPU-KRATINA et K.P. SRINATH<sup>1</sup>

## RÉSUMÉ

L'enquête sur l'emploi, la rémunération et les heures de travail que mène actuellement la Division du travail de Statistique Canada est une enquête mensuelle d'envergure permettant de recueillir des données à partir d'un grand échantillon d'établissements d'entreprises. Dans le présent article, l'auteur décrit les techniques d'enquête utilisées. Il donne une brève description des méthodes de stratification, de détermination de la taille de l'échantillon et de répartition, tandis qu'il décrit plus en détail la méthode de renouvellement en raison de sa complexité. Il souligne aussi certaines des simplifications qu'on peut apporter au plan de sondage.

**MOTS CLÉS:** Etablissement; fardeau de réponse; base de sondage.

## 1. INTRODUCTION

### 1.0 Objectifs de l'enquête

L'enquête sur l'emploi, la rémunération et les heures de travail (EBRH) est une enquête mensuelle menée auprès des établissements par Statistique Canada. Les principaux objectifs de l'EBRH sont les suivants:

- i) établir des estimations mensuelles du nombre total d'employés rémunérés, de la rémunération hebdomadaire moyenne, de la rémunération horaire moyenne, de la moyenne des heures de travail par semaine et d'autres variables connexes pour chaque division d'activités économiques au niveau provincial;
- ii) fournir les estimations susmentionnées pour le Canada au niveau des catégories à trois chiffres de la Classification des activités économiques (CAE);
- iii) déterminer les erreurs types de toutes les estimations produites.

Elle a aussi pour objet de produire annuellement des estimations pour chaque catégorie à trois chiffres de la CAE au niveau provincial.

L'enquête englobe toutes les activités économiques, à l'exception des secteurs suivants: agriculture, pêche et trappage, services domestiques privés, organisations religieuses et services militaires.

Dans le présent article, l'auteur décrit la méthode de sélection et de renouvellement de l'échantillon ainsi que la méthode d'estimation adoptée pour l'enquête. Le chapitre 2 présente la méthode de sélection et de renouvellement de l'échantillon en détail, tandis que le chapitre 3 est consacré à la méthode d'estimation. On trouve dans l'annexe certains détails relatifs au chapitre 2. L'annexe présente aussi une autre forme des estimations mensuelles des unités actives.

Pour une description complète de la méthode utilisée pour l'EBRH, voir le document de référence [3].

<sup>1</sup> Ioana Schiopu-Kratina et K.P. Srinath, Division des méthodes d'enquêtes-entreprises, Statistique Canada, Ottawa, KIA 0T6.



- CRUPDAS, A.M., REID, N., et COX, D.R. (1989). A time series illustration of approximate conditional likelihood. *Biometrika*, 76, 231-237.
- CUNIA, T., et CHEVROU, R.B. (1969). Sampling with partial replacement on three or more occasions. *Forest Science*, 15, 204-224.
- FRASER, D.A.S. (1967). Data transformations and the linear model. *The Annals of Mathematical Statistics*, 38, 1456-1465.
- HARVEY, A.C., et PHILLIPS, G.D.A. (1979). Maximum likelihood estimates of regression models with autoregressive-moving average disturbances. *Biometrika*, 66, 49-58.
- KALBFLEISCH, J.D., et SPROTT, D.A. (1970). Application of likelihood methods to models involving large numbers of parameters. *Journal of the Royal Statistical Society, Series B*, 32, 175-208.
- KILPATRICK, D.J. (1981). Optimum allocation in stratified sampling of forest populations on successive occasions. *Forest Science*, 27, 730-738.
- PATTERSON, H.D. (1950). Sampling on successive occasions with partial replacement of units. *Journal of the Royal Statistical Society, Series B*, 12, 241-255.
- ROBERTS, G., RAO, J.N.K., et KUMAR, S. (1987). Logistic regression analysis of sample survey data. *Biometrika*, 74, 1-12.
- TUNNICLIFFE WILSON, G. (1989). On the use of marginal likelihood in time series model estimation. *Journal of the Royal Statistical Society, Series B*, 51, 15-27.



prendre la fonction de vraisemblance définie en (3), la multiplier par la distribution de  $\beta$ , puis intégrer le tout par rapport à  $\beta$  pour obtenir la fonction de vraisemblance pour les paramètres du modèle des coefficients aléatoires. On aurait ainsi des matrices de la même dimension que  $\Gamma$ .

Puisque  $S$  est connue, nous pouvons obtenir facilement une estimation d' $\Omega$ , la matrice de corrélation des erreurs d'enquête. Nous pouvons aussi obtenir une valeur estimée de  $\kappa = \sigma^2/\gamma^2$ . Les hypothèses qui sous-tendent la fonction de vraisemblance marginale définie en (17) nous obligent à supposer que  $e_j$  dans l'équation (16) est une variable aléatoire stationnaire. Par conséquent, la moyenne des éléments diagonaux de  $S$  donne une valeur estimée de  $\sigma^2$ . Si  $\gamma^2$  est la variance des moyennes  $\mu$ , alors la variation entre  $y_i$ ,  $i = 1, \dots, k$  donne une valeur estimée de  $\sigma^2 + \gamma^2$ . De ces deux valeurs estimées, nous pouvons déduire une valeur estimée pour  $\kappa$ . Suivant le modèle (16),  $X$  dans l'équation (17) est la matrice unité  $k \times k$  tandis que  $W$  est un vecteur colonne  $k \times 1$  formé de uns. La fonction de vraisemblance marginale ainsi obtenue est une pseudo fonction puisque certains des paramètres ont été remplacés par des valeurs estimées. Dans ce cas, nous pouvons déterminer la pseudo fonction de vraisemblance marginale pour les paramètres dans  $\Gamma$  (pseudo car  $\kappa$  et  $\Omega$  ont été remplacés par leurs estimations) en prenant l'équation (17) et en effectuant les substitutions appropriées. Les paramètres dans  $\Gamma$  sont les paramètres de corrélation du processus ARMA appliqué à  $\mu_i$ . Si  $k$ , le nombre de passages, est relativement élevé par rapport au nombre de paramètres dans  $\Gamma$ , les estimateurs fondés sur la fonction marginale et sur la fonction conditionnelle approximative devraient être semblables à l'estimateur du maximum de vraisemblance. Au point de vue du calcul, il semble que la fonction de vraisemblance intégrale qui utilise les modèles d'espace d'états décrits par Binder et Dick (1989a, section 3) soit la plus simple à appliquer dans les circonstances.

## REMERCIEMENTS

Cette étude a été rendue possible grâce à une subvention du Conseil de recherches en sciences naturelles et en génie du Canada. L'auteur aimerait remercier l'arbitre de ses observations sur une version préliminaire de l'article.

## BIBLIOGRAPHIE

- BELLHOUSE, D.R. (1978). Marginal Likelihoods for distributed lag models. *Statistische Hefte*, 19, 2-14.
- BELLHOUSE, D.R. (1989). Optimal estimation of linear functions of finite population means in rotation sampling. *Journal of Statistical Planning and Inference*, 21, 69-74.
- BELLHOUSE, D.R. (1990). On the equivalence of marginal and approximate conditional likelihoods for correlation parameters under a normal model. *Biometrika*, 77, 743-746.
- BINDER, D.A., et HIDIROGLOU, M.A. (1988). Sampling in time. *Handbook of Statistics, Volume 6 (Sampling)* (éds. P.R. Krishniah et C.R. Rao). Amsterdam: North-Holland, 187-211.
- BINDER, D.A., et DICK, J.P. (1989). Enquêtes répétées – modélisation et estimation. *Techniques d'enquête*, 15, 31-48.
- BINDER, D.A., et DICK, J.P. (1990). Méthode pour l'analyse des modèles ARMMI. *Techniques d'enquête*, 16, 251-265.
- BLIGHT, B.J.N., et SCOTT, A.J. (1973). A stochastic model for repeated surveys. *Journal of the Royal Statistical Society, Series B*, 35, 61-68.
- COCHRAN, W.G. (1977). *Sampling Techniques*. 3ième édition. New York: Wiley.
- COX, D.R., et REID, N. (1987). Parameter orthogonality and approximate conditional inference (avec discussion). *Journal of the Royal Statistical Society, Series B*, 49, 1-39.

(15) 
$$(\hat{y}_{it} - \mu_i) / ((\text{def}_{it})^{1/2} - \mu_{i-1} - \mu_i) = \phi (\hat{y}_{i-1} - \mu_{i-1})^{1/2} + \epsilon_i$$

où  $\epsilon_i$  a une variance constante. Cela cadre bien avec le modèle (2), où le vecteur d'observations  $y$  renferme des données de la forme  $\hat{y}_{it} / ((\text{def}_{it})^{1/2})$ ,  $\beta$  est  $(\mu_1, \mu_2, \dots, \mu_k)^T$  et  $X$  renferme des éléments de la forme  $1 / ((\text{def}_{i-1})^{1/2})$ . La fonction de vraisemblance marginale, considérée en l'occurrence comme un cas particulier des équations (5) ou (6), peut être évaluée à l'aide du modèle d'espace d'états de Harvey et Phillips (1979) et dont nous avons fait mention dans la section 2. Compte tenu du modèle défini ci-dessus (équ. 20 et 21), il est souhaitable de recourir à l'estimation fondée sur la fonction de vraisemblance marginale et la fonction de vraisemblance conditionnelle approximative. La valeur estimée de  $\phi$  repose en l'occurrence sur la variation entre les estimations élémentaires dans chaque groupe de renouvellement, la variance à l'intérieur de ces estimations n'étant pas connue. Comme un groupe de renouvellement passe relativement peu de temps dans l'échantillon, il y a de fortes chances que les estimateurs du maximum de vraisemblance soient biaisés et non convergents.

5. ANALYSE

Binder et Dick (1990) ont aussi proposé d'utiliser des techniques d'estimation fondée sur la fonction de vraisemblance marginale pour l'échantillonnage répété. Suivant leur cadre, nous supposons que nous connaissons les valeurs estimées des moyennes,  $\hat{y}_i$ , pour chaque passage  $t = 1, \dots, k$ . Supposons aussi que la matrice,  $S$ , des variances-covariances des estimations est connue. Comme dans Binder et Dick (1989, 1990) notamment, les  $\hat{y}_i$  peuvent être définies par le modèle

(16) 
$$\hat{y}_i = \mu_i + \epsilon_i,$$

où  $\epsilon_i$  est l'erreur d'enquête au passage  $i$ , la matrice des variances-covariances estimée étant représentée par  $S$ . Les moyennes pour chaque passage ( $\mu_i$  pour le passage  $i$ ) suivent un processus ARMA. Comme le modèle (16) est un cas particulier du modèle de régression avec coefficients aléatoires, la fonction de vraisemblance marginale appropriée est différente de l'équation (5).

On obtient une fonction de vraisemblance marginale ou une fonction de vraisemblance conditionnelle approximative pour les paramètres de corrélation dans un modèle de régression avec coefficients aléatoires de la façon suivante. Supposons, pour ce qui a trait au modèle (2), que  $\beta$  est un vecteur aléatoire défini par l'équation  $\beta = W\delta + u$ , où  $W$  est une matrice  $p \times q$  de valeurs connues,  $\delta$  est un vecteur de paramètres de dimensions  $q \times 1$  et  $u \sim N(0, \gamma^2 I)$ , sans lien avec  $\epsilon$ . Selon le modèle composé  $y = XW\delta + Xu + \epsilon$ , la fonction de vraisemblance logarithmique pour  $\delta, W, \Gamma, \gamma^2$ , et  $\kappa = \sigma^2 / \gamma^2$ , désignée par  $L(\delta, \kappa, \gamma^2, \Gamma, W)$ , est définie par l'équation (3), où  $W$  est remplacé par l'expression  $\kappa W + X\Gamma X^T$  et  $X\beta$  par  $XW\delta$ . De la même façon, la fonction de vraisemblance marginale, désignée par  $L_M(\kappa, \Gamma, W)$ , est définie par l'équation (5), où  $X$  est remplacé par  $XW$  et  $W$  par  $\kappa W + X\Gamma X^T$ . Ainsi,

ou 
$$L_M(\kappa, \Gamma, W) = \{ | \kappa W + X\Gamma X^T |^{1/2} | (XW)^T (\kappa W + X\Gamma X^T)^{-1} XW |^{1/2} g_{m-b}^{-1} \} \quad (17)$$
$$g = y^T (\kappa W + X\Gamma X^T)^{-1} y$$

Or, la dimension de  $W$  peut être plus grande par rapport à celle de  $\Gamma$ , c'est une situation qu'on observe parfois dans l'échantillonnage répété. Pour redresser cette situation, on pourrait



En substituant ces données dans l'équation (13), il est possible d'obtenir la fonction de vraisemblance marginale et la fonction de vraisemblance conditionnelle approximative correspondant aux données pour le paramètre autotégressif du premier degré  $\phi$ . Voir la figure 1.

#### 4. ENQUÊTES À PLAN DE SONDAGE COMPLEXE

Il y a plusieurs façons d'analyser des données de séries chronologiques tirées d'enquêtes à plan de sondage complexe. Chaque méthode qu'on peut proposer dépendra des données d'échantillon qui auront pu être recueillies.

Si, par exemple, on dispose de microdonnées, il est possible de calculer pour chaque groupe de renouvellement la matrice, fondée sur le plan de sondage complexe, des variances-covariances des estimations élémentaires. On définit une pseudo fonction de vraisemblance marginale en remplaçant  $\bar{x}_i$  et  $S_i$  dans les équations (5) et (8) par leurs équivalents pour les enquêtes à plan de sondage complexe. C'est la méthode qu'utilisent par exemple Roberts, Rao et Kumar (1987) dans l'analyse de régression logistique pour plans de sondage complexes: déterminer une fonction de vraisemblance ou un ensemble d'équations de vraisemblance et remplacer les paramètres statistiques habituels par leurs équivalents pour les enquêtes à plan de sondage complexe.

Suivant un plan d'échantillonnage aléatoire simple, la matrice  $S_i$  estime la matrice des variances-covariances de la population finie pour les valeurs observées aux passages incluant le groupe de renouvellement  $i$ . Ainsi, dans un plan de sondage complexe, on remplace la matrice  $S_i$  par un estimateur, consistant avec le plan, de la matrice des variances-covariances de la population finie correspondante. Par exemple, Kilpatrick (1981) a examiné un plan de sondage stratifié à deux passages pour évaluer le nombre d'arbres sur pied dans les forêts appartenant à l'Etat en Irlande du Nord; les strates étaient fondées sur les époques, en commençant par les années vingt, où les forêts ont été plantées. Afin de calculer l'équivalent de  $S_i$  dans un plan de sondage stratifié, il est nécessaire d'avoir l'estimateur des moyennes à chaque passage, les moyennes des strates, les variances des strates ainsi que les covariances des strates pour les unités qui étaient présentes et non présentes dans l'échantillon aux deux passages. Pour une population stratifiée, on peut décomposer la variance (ou covariance) de la population finie en des termes comprenant la variation (ou la covariance) entre les strates et la variation (ou la covariance) à l'intérieur des strates; voir, par exemple, Cochran (1977, équation 5.32). On utiliserait les valeurs estimées des moyennes et les moyennes des strates pour obtenir des estimations convergentes de la variation ou de la covariance entre les strates, et les valeurs estimées des variances et des covariances des strates pour obtenir des estimations de la variation et de la covariance à l'intérieur des strates. Malheureusement, seuls certains estimateurs des variances et des covariances des strates s'appliquaient à l'étude de Kilpatrick, de sorte que l'article ne contient pas suffisamment de données pour pouvoir calculer un estimateur du maximum de vraisemblance marginal pour la corrélation entre les volumes de bois aux deux passages.

Or, il arrive rarement que l'on dispose de microdonnées. La méthode d'estimation dépend alors des données disponibles. Nous envisageons ici un scénario mais d'autres sont possibles. Supposons que nous connaissions seulement les estimations élémentaires et les effets du plan correspondant. Soit  $y_{it,r}$  l'estimation tirée du groupe de renouvellement  $G_i$  au passage  $t$  et fondée sur un échantillon de taille  $m_r$ . Soit  $\text{def}_{it,r}$  l'effet du plan qui correspond à  $y_{it,r}$ . Si  $\sigma^2/m_r$  est la variance de  $y_{it,r}$  suivant un échantillonnage aléatoire simple, alors, en vertu du théorème central limite,

$$(y_{it,r} - \mu_r) / (\text{def}_{it,r})^{1/2} \sim N(0, \sigma^2/m_r) \tag{14}$$

approximativement. Nous pouvons établir un modèle en supposant, pour  $G_i$ , un processus ARMA comme celui ci-dessous



pour  $t = 1, \dots, k$ , où  $\pi_1 = \pi_{k+1} = 0$  et  $y_1'' = y_1'$ . Le vecteur des moyennes estimées  $\hat{\mu}$  est non biaisé pour  $\mu$  selon le modèle (9) et sa matrice des variances-covariances correspondante est  $\sigma^2 G^{-1}$ . Ainsi, d'après les équations (5) ou (6), la fonction de vraisemblance marginale et la fonction de vraisemblance conditionnelle approximative pour  $\phi$  est

$$L_M(\phi) = \frac{\{A(\hat{\mu}, \phi) + B(\phi)\}^{(m-k)/2} |G|^{1/2}}{(1 - \phi^2)^{d/2}} \quad (13)$$

### 3.3 Exemple

Les données utilisées aux fins du présent exemple sont des données sur laylviculture tirées de Cunia et Chevrou (1969, p. 220). Elles représentent le volume négociable de bois de cons-truction par tracé observé à trois passages avec remise partielle des unités d'échantillon. Dans un échantillonnage avec renouvellement, on suppose que, une fois qu'on laisse tomber une unité de l'échantillon, celle-ci n'est pas sélectionnée de nouveau. Compte tenu de cette hypo-thèse, on a rectifié les données dans Cunia et Chevrou. En particulier, on a laissé tomber du présent exemple les valeurs observées pour les unités qui étaient présentes dans l'échantillon aux premier et troisième passages, mais qui n'étaient pas présentes dans l'échantillon au deuxième passage. À partir des données restantes, il est possible de faire les calculs suivants:

$\pi_2 = 86/139$ ,  $\pi_3 = 76/100$ ,  $n_1 = 104$ ,  $n_2 = 139$ ,  $n_3 = 100$ ,  $y_2' = 161.5581$ ,  $y_3' = 179.9211$ ,  $y_1'' = 154.0673$ ,  $y_2'' = 167.2075$ ,  $y_3'' = 181.125$ ,  $x_1' = 147.6512$ ,  $x_2' = 163.4342$ ,  $xy_2' = 864129.2$ ,  $xy_3' = 555369.5$ ,  $xy_1'' = 943948.5$ ,  $xy_2'' = 266820.7$ ,  $xy_3'' = 271762.6$ ,  $sxx_1' = 800753.5$ ,  $sxx_2' = 559850.7$ ,  $sxx_3' = 812435.7$ ,  $sxy_3' = 550943.6$ ,  $d = 181$ , et  $m - k = 340$ .

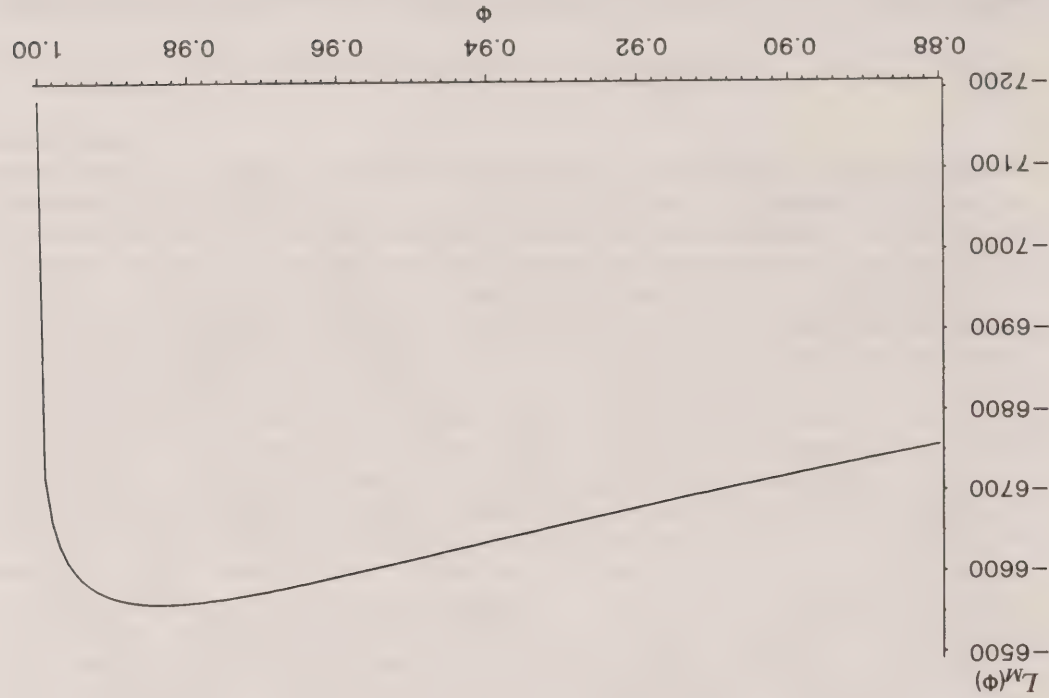


Figure 1. Fonction de vraisemblance marginale pour le paramètre AR(1)

$sy'_t$  = la somme des carrés pour les unités échantillonnées au passage  $t$ , qui étaient aussi présentes dans l'échantillon au passage précédent ( $t - 1$ );

$sy''_t$  = la somme des carrés pour les unités échantillonnées au passage  $t$  et qui n'étaient pas présentes dans l'échantillon au passage précédent ( $t - 1$ );

$sxx'_t$  = la somme des carrés pour les unités échantillonnées au passage  $t$ , qui sont aussi présentes dans l'échantillon au passage suivant ( $t + 1$ );

$sy_t$  = la somme des carrés pour toutes les unités échantillonnées au passage  $t$ ;

$sxy'_t$  = la somme des produits pour les observations relatives aux unités échantillonnées au passage  $t$ , qui étaient aussi présentes dans l'échantillon au passage précédent ( $t - 1$ ).

Suivant le cas particulier du modèle (9), et après de nombreuses transformations algébriques, nous pouvons montrer que, lorsqu'on fait la sommation de l'expression (7) pour tous les groupes de renouvellement  $r$ , la fonction de vraisemblance logarithmique pour les données se ramène à

$$L(\mu_1, \dots, \mu_k, \sigma^2, \phi) = -m \ln \sigma + (d/2) \ln(1 - \phi^2)$$

$$- \{A(\mu, \phi) + B(\phi)\} / (2\sigma^2), \quad (10)$$

où  $d$  est le nombre d'unités distinctes échantillonnées (c'est-à-dire abstraction faite du nombre de fois qu'une unité est échantillonnée) et  $m$  est la taille de l'échantillon global ( $n_1 + \dots + n_k$ ).

De plus, dans l'équation (10),

$$A(\mu, \phi) = (1 - \phi^2)n_1(y_1 - \mu_1)^2$$

$$+ \sum_{k=2}^t [x'_t n_t \{y'_t - \mu_t - \phi(x'_{t-1} - \mu_{t-1})\}^2 + (1 - \pi_t)n_t(1 - \phi^2)(y''_t - \mu_t)^2] \quad (11)$$

et

$$B(\phi) = (1 - \phi^2)sy_t + \sum_{k=2}^t \{\phi^2 sxx'_{t-1} - 2\phi sxy'_t + syy'_t + (1 - \phi^2)syy''_t\}. \quad (12)$$

Pour n'importe quelle valeur donnée de  $\phi$ , les estimateurs du maximum de vraisemblance sont  $\hat{\mu} = G^{-1}z$  et  $\hat{\sigma}^2 = \{A(\hat{\mu}, \phi) + B(\phi)\}/m$ , où  $A(\hat{\mu}, \phi)$  est défini par l'équation (11),  $\mu$  étant remplacé par son estimateur du maximum de vraisemblance, et où  $G$  est une matrice bande symétrique  $k \times k$  de largeur de bande 3 et  $z$  est un vecteur  $k \times 1$ . Les éléments non nuls de  $G$  sont définis

$$g_{tt} = \pi_t n_t + (1 - \pi_t)n_t(1 - \phi^2) + \pi_{t+1}n_{t+1}\phi^2, \quad \text{pour } t = 1, \dots, k$$

et

$$g_{t,t+1} = -\pi_{t+1}n_{t+1}\phi, \quad \text{pour } t = 1, \dots, k - 1,$$

où  $\pi_1 = \pi_{k+1} = 0$ . Les éléments de  $z$  sont définis

$$z_t = \pi_t n_t (y'_t - \phi x'_{t-1}) + (1 - \pi_t)n_t y''_t(1 - \phi^2) - \pi_{t+1}n_{t+1}(y'_{t+1} - \phi x'_t),$$

Etant donné les paramètres dans  $\Omega$ , ou bien les paramètres du maximum de vraisemblance dans  $\Omega_1, \dots, \Omega^{k+c-1}$ , il est possible de déterminer des expressions pour les estimateurs du maximum de vraisemblance  $\hat{\mu}$  et  $\hat{\sigma}^2$ , qui servent à estimer  $\mu$  et  $\sigma^2$  respectivement. De même, il est possible de connaître la matrice des variances-covariances estimée de  $\hat{\mu}$ ,  $V(\hat{\mu})$ . C'est ce que nous illustrons pour un processus autorégressif du premier degré dans la sous-section 3.2. La fonction de vraisemblance marginale pour  $\Omega_1, \dots, \Omega^{k+c-1}$  est alors définie par l'équation (5), où

$$|\Phi|^{1/2} = \prod_{r=1}^{k+c-1} |\Omega_r|$$

$$|X^T \Phi^{-1} X|^{1/2} = V(\hat{\mu})/s^k, \quad (8)$$

$$s^2 = \sum_{r=1}^{k+c-1} \{n_r x_r^T \Omega_r^{-1} x_r + (n_r - 1) \text{tr}(\Omega_r^{-1} S_r)\},$$

et  $p = k$ ; dans l'équation ci-dessus,  $x_r$  est  $x_p$  à la différence près que  $\mu$  est remplacée par l'estimateur du maximum de vraisemblance correspondant.

### 3.2 Processus autorégressifs du premier degré

Lorsqu'on utilise des formes précises des matrices de corrélation  $\Omega_1, \dots, \Omega^{k+c-1}$ , il est possible de simplifier quelque peu la forme générale de la fonction de vraisemblance marginale pour les paramètres de corrélation, définie en (5) et en (6). Par exemple, supposons le modèle autorégressif du premier degré

$$y_{it} = \mu + \phi (y_{i,t-1} - \mu_{t-1}) + \epsilon_{it}, \quad (9)$$

où  $\epsilon_{it} \sim N(0, \sigma^2)$  pour  $t = 1, \dots, k$  et  $j = 1, \dots, N$ , et où les  $\epsilon$  sont mutuellement indépendants. Le modèle (9), qui correspond essentiellement au modèle de Patterson (1950), est un cas particulier du modèle (1). Comme dans la sous-section 3.1, le vecteur des paramètres de régression  $\beta$  est  $(\mu_1, \dots, \mu_k)^T$ . Lorsque le vecteur de données  $y$  contient pour chaque unité les observations groupées selon les passages où cette unité a été échantillonnée, conformément à la description qui est faite de l'échantillonnage avec renouvellement dans la sous-section 3.1, on peut exprimer la matrice de corrélation  $\Phi$ , qui est désormais une fonction du paramètre autorégressif  $\phi$ , comme une somme directe de matrices qui sont toutes les matrices de corrélation d'un processus autorégressif du premier degré.

Nous reprenons la notation utilisée par Patterson (1950) pour désigner des tailles d'échantillon, des moyennes ainsi que des sommes des carrés et des produits (redressées en fonction de la moyenne pertinente) pour le passage  $t$ :

$n_t$  = le nombre d'unités échantillonnées au passage  $t$ ;  
 $y'_t$  = la moyenne pour les unités échantillonnées au passage  $t$ , qui étaient aussi présentes dans l'échantillon au passage précédent ( $t - 1$ );  
 $y''_t$  = la moyenne pour les unités échantillonnées au passage  $t$  et qui n'étaient pas présentes dans l'échantillon au passage précédent ( $t - 1$ );  
 $\bar{y}_t$  = la moyenne pour toutes les unités échantillonnées au passage  $t$ ;  
 $x'_t$  = la moyenne pour les unités échantillonnées au passage  $t$ , qui sont aussi présentes dans l'échantillon au passage suivant ( $t + 1$ );



$$L(\theta, \lambda) \mid I(\theta, \lambda) \mid_{1/2},$$

où  $I(\theta, \lambda)$  est la matrice d'information observée pour  $\lambda$  à une valeur déterminée de  $\theta$ . Voir Cox et Reid (1987, équation 10).

À la suite de Cruddas et coll. (1989), Bellhouse (1990) a proposé pour le modèle (2) la transformation de paramètre  $\lambda = \ln \sigma + (\ln |\Omega|)/(2m)$ ,  $\beta$  demeurant inchangé. Suivant ces nouvelles conditions, la fonction de vraisemblance logarithmique est désignée par  $L(\beta, \lambda, \Phi)$  et peut être tirée de l'équation (3). Si les éléments de  $\Phi$  sont des fonctions d'un paramètre  $\phi$ , les paramètres dérangeants  $\lambda$  et  $\beta$  sont l'un et l'autre orthogonaux à  $\phi$ , c'est-à-dire

$$-\frac{1}{m} E \left[ \frac{\partial \phi \partial \lambda}{\partial^2 L(\beta, \lambda, \Phi)} \right] = 0$$

et

$$-\frac{1}{m} E \left[ \frac{\partial \phi \partial \beta}{\partial^2 L(\beta, \lambda, \Phi)} \right] = 0,$$

lorsque chaque élément de  $\Phi$  est une fonction continue et différentiable de  $\phi$ . En outre dans ce cas, la fonction conditionnelle approximative pour  $\Phi$ ,  $L_C(\Phi)$  est identique à la fonction marginale,  $L^M(\Phi)$ , définie en (5) ou en (6). Voir Bellhouse (1990) pour plus de détails.

Il est possible d'évaluer la FM et la FCA définies en (5) ou en (6) pour n'importe quelle matrice  $\Phi$  en se servant de modèles d'états à la manière de Harvey et Phillips (1979). Une fois que les calculs récursifs visant à estimer  $\beta$  et  $\sigma^2$  sont terminés, on peut calculer, pour n'importe quelle matrice  $\Phi$ , la valeur de  $s^2$  et de  $|\Phi|_{1/2}$  au moyen des formules proposées par Harvey et Phillips (1979, équations 5.6 et 6.6, et 4.3 respectivement). On n'a alors qu'à calculer  $X^T \Phi^{-1} X$  et son déterminant. La dernière étape du processus récursif de Harvey et Phillips (1979, équation 3.4) permet de déterminer la valeur de  $X^T \Phi^{-1} X$ .

### 3. ÉCHANTILLONNAGE ALÉATOIRE SIMPLE RÉPÉTÉ

#### 3.1 Quelques résultats pour l'échantillonnage avec renouvellement

Supposons que le groupe de renouvellement  $G_r$  est choisi la première fois au passage  $u$  et la dernière fois au passage  $v$ ,  $u$  étant égal à 1 ou à  $r$  et  $v$  correspondant à  $r + c - 1$  ou à  $k$ . Le nombre total de passages où une unité de  $G_r$  est incluse dans l'échantillon est  $b = v + 1 - u$ . Soient  $y_{u,r}, \dots, y_{v,r}$  les moyennes d'échantillon ou les estimations élémentaires pour  $G_r$  aux passages  $u, u + 1, \dots, v - 1, v$  respectivement. Alors, suivant le modèle (1), la contribution de  $G_r$  à la fonction de vraisemblance logarithmique définie en (3) est

$$- \left\{ b n_r \ln \sigma + (n_r/2) \ln(|\Omega_r|) + [n_r x_r^T \Omega_r^{-1} x_r + (n_r - 1) \text{tr}(\Omega_r^{-1} S_r)] / (2\sigma^2) \right\}, \quad (7)$$

où  $x_r^T$  est le vecteur  $1 \times b$  ( $y_{u,r}, \dots, y_{v-1,r}, \mu_{u+1}, \dots, y_{v-1}, r - \mu_{v-1}, y_{v,r}, \mu_v$ ),  $S_r$  est la matrice  $b \times b$  des variances et des covariances de l'échantillon pour les observations relatives au groupe de renouvellement et où  $\Omega_r$  est la matrice de corrélation  $b \times b$  des observations relatives à une unité du groupe de renouvellement. Notons que les paramètres dans  $\Omega$  seront aussi les paramètres dans  $\Omega_r$ . La matrice de corrélation  $\Omega$  est fondée sur les valeurs observées à tous les passages 1 à  $k$ ; la matrice de corrélation  $\Omega_r$  est tirée du sous-ensemble de données observées aux passages  $u$  à  $v$ . En vertu de l'hypothèse d'indépendance, on obtient la fonction de vraisemblance logarithmique en faisant la sommation de l'expression (7) pour tous les groupes de renouvellement.

où le vecteur d'erreurs  $\epsilon \sim N(0, \sigma^2 \Phi)$ , étant la matrice de corrélation de dimensions  $m \times m$  et  $\beta$  le vecteur des coefficients de régression  $p \times 1$  de sorte que  $X$  est de dimensions  $m \times p$ . La fonction de vraisemblance logarithmique pour  $\beta$ ,  $\sigma^2$  et  $\Phi$  est définie

$$L(\beta, \sigma^2, \Phi) = - \{ m \ln \sigma + (\ln |\Phi| + (y - X\beta)^T \Phi^{-1} (y - X\beta)) / (2\sigma^2) \}. \quad (3)$$

Pour une valeur donnée de  $\Phi$ ,

$$\hat{\beta} = (X^T \Phi^{-1} X)^{-1} X^T \Phi^{-1} y$$

et

$$s^2 = y^T \Phi^{-1} y - y^T \Phi^{-1} X (X^T \Phi^{-1} X)^{-1} X^T \Phi^{-1} y$$

sont des estimateurs exhaustifs conjoints de  $\beta$  et  $\sigma^2$ .

On obtient une fonction de vraisemblance marginale pour  $\Phi$  par une réduction des données  $y$  aux statistiques exhaustives  $\hat{\beta}$  et  $s^2$  et à la statistique ancillaire

$$u = \Phi^{-1/2} (y - X (X^T \Phi^{-1} X)^{-1} X^T \Phi^{-1} y) / s,$$

où  $\Phi^{-1/2}$  est la matrice de dimensions  $m \times m$  telle que  $\Phi^{-1} = \Phi^{-1/2} \Phi^{-1/2}$ . La fonction de vraisemblance marginale de  $\Phi$  correspond à la distribution marginale de la statistique ancillaire  $u$  multipliée par le produit des différentielles  $da_i$ ,  $i = 1, \dots, m$ . Voir Kalbfleisch et Sprott (1970, équations 6 et 10) pour une analyse globale et une expression générale pour  $Ida_i$ . Bellhouse (1978), suivi quelques années plus tard de Tunniciiffe Wilson (1989), a montré que la fonction de vraisemblance marginale pour  $\Phi$  suivant un modèle normal était définie

$$L_M(\Phi) = \{ |\Phi|^{1/2} | X^T \Phi^{-1} X |^{1/2} s^{m-p} \}^{-1}. \quad (5)$$

Notons que l'équation (4) est proportionnelle à l'estimateur du maximum de vraisemblance de  $\sigma^2$  étant donné  $\Phi$  et que  $s^2 (X^T \Phi^{-1} X)^{-1}$  est proportionnelle à la matrice des variances-covariances estimée de l'estimateur du maximum de vraisemblance de  $\beta$  étant donné  $\Phi$ . Alors, l'équation (5) peut être réécrite

$$L_M(\Phi) = \frac{s^m |\Phi|^{1/2}}{|\text{est var}(\hat{\beta})|^{1/2}}. \quad (6)$$

Pour définir une fonction de vraisemblance conditionnelle approximative (FCA), il faut d'abord transformer les paramètres de manière à obtenir une relation d'orthogonalité entre les paramètres d'intérêt et les paramètres dérangeants, qui peuvent dépendre des premiers. Il y a orthogonalité entre les ensembles de paramètres lorsque la matrice d'information correspondante est une matrice diagonale en blocs, chaque bloc servant lui-même de matrice d'information pour un ensemble de paramètres. La fonction de vraisemblance conditionnelle est liée à la distribution des données  $y$ , qui dépend elle-même de l'estimateur du maximum de vraisemblance des paramètres dérangeants pour les valeurs déterminées des paramètres d'intérêt. On obtient la fonction de vraisemblance conditionnelle approximative en appliquant deux approximations à cette distribution conditionnelle. Voir Cox et Reid (1987, sous-section 4.1), pour une analyse des calculs. Par exemple, posons  $\Theta$  comme le vecteur des paramètres d'intérêt et  $\Lambda$ , qui dépend vraisemblablement de  $\Theta$ , comme le vecteur des paramètres dérangeants orthogonal à  $\Theta$ . La fonction de vraisemblance complète pour les paramètres  $\Theta$  et  $\Lambda$  est désignée par  $L(\Theta, \Lambda)$  tandis que la fonction profil pour  $\Theta$ ,  $L(\Theta, \hat{\Lambda})$  équivaut à la fonction de vraisemblance ordinale à la différence près que  $\hat{\Lambda}$  est remplacé par l'estimateur du maximum de vraisemblance correspondant. La fonction de vraisemblance conditionnelle approximative pour  $\Theta$  est

Lorsque la valeur  $c$  est peu élevée, les valeurs estimées pour les paramètres de corrélation dans  $\Omega$  peuvent alors être instables, causant une instabilité dans l'estimateur d'intérêt  $(\mu_1, \mu_2, \dots, \mu_k)^T$ . Considère d'une autre manière, le nombre total de paramètres est d'au moins  $k + 2$  et augmentera avec le temps, c'est-à-dire à chaque nouveau passage. Étant donné que la dimension de l'espace des paramètres augmentera avec le temps, l'estimateur du maximum de vraisemblance pourrait être biaisé et non convergent. Le problème de la stabilité des estimateurs dans un échantillonnage répété a été étudié, par exemple, par Blight et Scott (1973), qui supposent que les éléments de  $(\mu_1, \mu_2, \dots, \mu_k)^T$  suivent un processus de série chronologique. À partir de cette hypothèse, on fixe la dimension de l'espace des paramètres à un nombre relativement peu élevé de sorte qu'on règle les problèmes d'instabilité, de biais et de non convergence. Dans le présent article, l'auteur adopte une approche différente. Il conserve l'hypothèse des moyennes fixes et établit la fonction de vraisemblance marginale (FM) et la fonction de vraisemblance conditionnelle approximative (FCA) pour les paramètres dans  $\Omega$ , traitant les moyennes fixes comme des paramètres dérangeants.

On a proposé pour la première fois les FM comme une méthode générale pour éliminer les paramètres dérangeants de la fonction de vraisemblance (Fraser 1967; Kalbfleisch et Sprott 1970). Les FCA ont été définies dans le même but par Cox et Reid (1987). Ceux-ci affirment que la FCA est préférable à la fonction profil de vraisemblance, que l'on obtient en remplaçant les paramètres dérangeants dans la fonction de vraisemblance par l'estimation la plus vraisemblable correspondante lorsque les paramètres d'intérêt sont connus. Bellhouse (1990) a démontré l'équivalence de la FM et de la FCA pour des paramètres de corrélation suivant un modèle normal. S'inspirant de l'étude de Cox et Reid, Cruddas et coll. (1989) ont établi une FCA pour les paramètres de corrélation dans plusieurs petites séries de processus autorégressifs du premier degré ayant la même variance et les mêmes paramètres d'autocorrélation. Ils ont montré par une étude de simulation que l'estimateur fondé sur la FCA était beaucoup moins biaisé que l'estimateur du maximum de vraisemblance fondé sur la fonction profil et qu'il recouvrait mieux l'intervalle de confiance. La situation décrite par Cruddas et ses coll. (1989) s'applique directement aux enquêtes à passages répétés. Afin de réduire le fardeau de réponse des personnes qui participent à ce genre d'enquêtes, on fait en sorte qu'elles ne fassent pas trop longtemps partie de l'échantillon. On prévoit que l'utilisation des FM et des FCA permettra d'obtenir de meilleures valeurs estimées des paramètres de corrélation et, par conséquent, de meilleures valeurs estimées de la moyenne pour chaque passage.

À l'intérieur d'un groupe de renouvellement, les données sur chaque personne sont normalement modélisées à l'aide d'un processus autorégressif de moyennes mobiles (ARMA), c'est-à-dire que les paramètres dans  $\Omega$  sont composés des paramètres de corrélation du processus ARMA. Voir Binder et Hidiroglou (1988) pour une étude de l'application des modèles de séries chronologiques à l'échantillonnage répété. Par conséquent, il est utile de définir des FM et des FCA suivant des modèles ARMA appliqués à un échantillonnage avec renouvellement. Dans la section 2, nous déterminons les FM et les FCA pour les paramètres de corrélation suivant un modèle normal. Nous appliquons ensuite les résultats de cette section aux enquêtes à passages répétés avec plan d'échantillonnage aléatoire simple d'unités dans des groupes de renouvellement. Enfin, dans la section 4, nous présentons plusieurs méthodes qui permettent d'appliquer ces fonctions de vraisemblance à des plans de sondage complexes.

## 2. FONCTION DE VRAISEMBLANCE MARGINALE ET FONCTION DE VRAISEMBLANCE CONDITIONNELLE APPROXIMATIVE POUR DES PARAMÈTRES DE CORRÉLATION SUIVANT UN MODÈLE NORMAL

Supposons que  $y$  est un vecteur d'observations de dimensions  $m \times 1$  suivant le modèle linéaire

$$y = X\beta + \epsilon$$

(2)



# Fonctions de vraisemblance marginales et fonctions de vraisemblance conditionnelles approximatives pour l'échantillonnage répété

D.R. BELLHOUSE<sup>1</sup>

## RÉSUMÉ

L'auteur définit des fonctions de vraisemblance marginales et des fonctions de vraisemblance conditionnelles approximatives pour les paramètres de corrélation d'un modèle de régression linéaire normal à erreurs corrélées. L'auteur se sert du principe de vraisemblance pour déterminer des fonctions de vraisemblance marginales et des fonctions de vraisemblance conditionnelles approximatives pour les paramètres de corrélation dans un plan d'échantillonnage répété (échantillonnage aléatoire simple et plans plus complexes).

MOTS CLÉS: Inférence fondée sur la vraisemblance; échantillonnage dans le temps; modèles ARMA; modèles d'états.

## 1. INTRODUCTION

Considérons une population finie de  $N$  unités qui peut être sondée  $k$  fois. Désignons par  $y_{jt}$  la valeur observée pour l'unité de population  $j$  au  $t$ -ième passage de l'enquête,  $j = 1, \dots, N$  et  $t = 1, \dots, k$ . Nous supposons que les unités de population sont indépendantes les unes des autres mais qu'il y a corrélation dans le temps pour les valeurs observées pour la même unité. En particulier, nous supposons que pour  $j$  importe quelle unité  $j$ ,

$$(1) \quad (y_{1j}, y_{2j}, \dots, y_{kj})^T \sim N(\mu, \sigma^2 \Omega),$$

où  $\Omega$  est une matrice de corrélation  $k \times k$  et  $\mu$  est le vecteur  $1 \times k$  de moyennes fixes  $(\mu_1, \mu_2, \dots, \mu_k)^T$ . En vertu de l'hypothèse du modèle explicite en (1), l'auteur utilise dans le présent article une méthode d'estimation fondée sur des modèles. À partir de tous les échantillons prélevés aux  $k$  passages, il est utile d'estimer  $(\mu_1, \mu_2, \dots, \mu_k)^T$ . La forme d'estimation fondée sur un modèle  $(\hat{\mu}_1, \hat{\mu}_2, \dots, \hat{\mu}_k)^T$ , lorsqu'elle est définie par la méthode du maximum de vraisemblance ou par celle des moindres carrés généralisés, par exemple, dépendra de  $\sigma^2$  et des paramètres dans  $\Omega$ . Par conséquent, nous devons obtenir des estimateurs valables pour  $\sigma^2$  et les paramètres dans  $\Omega$ .

Nous reprenons la notation de Bellhouse (1989) pour décrire le plan d'échantillonnage consistant à déter  $c$  groupes de renouvellement. Le groupe de renouvellement  $r$  ( $r = 1, 2, \dots, k + c - 1$ ), désigné par  $G_r$ , contient  $m_r$  unités. Au passage  $t$  ( $t = 1, \dots, k$ ), l'échantillon comprend les unités des groupes  $G_r, G_{r+1}, \dots, G_{r+c-1}$ , de sorte que sa taille  $n_t$  est égale à  $m_r + m_{r+1} + \dots + m_{r+c-1}$ . Au passage  $t + 1$ , on laisse tomber  $G_r$  de l'échantillon et on ajoute  $G_{r+c}$ . Chaque groupe de renouvellement est choisi sans remise parmi les unités de population qui n'ont pas été prélevées auparavant. Pour les  $k$  passages considérés globalement, la taille de l'échantillon est  $m = n_1 + n_2 + \dots + n_k$ . Le nombre maximal de passages où une unité reste dans l'échantillon est  $c$ .

<sup>1</sup> D.R. Bellhouse, Department of Statistical and Actuarial Sciences, University of Western Ontario, London, Ontario, Canada N6A 5B9.



- LAFLAMME, F. (1990). Étude comparative entre trois différentes populations visées par l'EPA selon leur type de service téléphonique. Document interne, Division des méthodes d'enquêtes sociales, Statistique Canada.
- MIAN, I.U.H. (1990). Dual frame estimation of proportions in sample surveys. Document interne, Division des méthodes d'enquêtes sociales, Statistique Canada.
- MUIRHEAD, R.C., GOWER, A.R., et NEWTON, F.T. (1975). The telephone experiment in the Canadian Labour Force Survey. *Survey Methodology*, 1, 158-180.
- TREWIN, D., et LEE, H. (1988). International comparisons of telephone coverage. *Telephone Survey Methodology*, (éds. R. Groves et coll.), New York: Wiley, 9-24.
- WAKSBERG, J. (1978). Sampling methods for random digit dialing. *Journal of the American Statistical Association*, 73, 576-579.



## 8. RÉSUMÉ

La méthodologie de collecte des données actuelle de la population active comprend une interview personnelle pour les ménages pendant leur premier mois dans l'échantillon et principalement des interviews téléphoniques au cours des mois suivants. La mise en oeuvre des interviews téléphoniques au cours des mois suivants a eu lieu au cours des années 70 pour les grandes régions urbaines et au cours des années 80 dans les autres régions. Dans les deux cas, les interviews téléphoniques se sont traduites par des économies appréciables des coûts sans aucune incidence sur les taux de réponse ou les estimations d'enquête. Avant le début du programme de recherche et de tests téléphoniques et IAO, en 1985, 80% des interviews EPA se faisaient par téléphone. Cette proportion a légèrement augmenté pour passer à 83% grâce à la mise en oeuvre du suivi téléphonique pour les ménages qui n'ont pu être contactés pendant une visite personnelle initiale et par la fourniture aux intervieweurs des numéros de téléphone de l'échantillon des appartements.

Le principal avantage du programme de recherche et d'essais a été de déterminer les options de la base de sondage et de la collecte des données à conserver pendant le remaniement post-censitaire de 1991 de l'enquête, ce qui comprend le maintien de l'organisation locale actuelle des intervieweurs, l'adoption des interviews personnelles assistées par ordinateur, le maintien des approches de la base de sondage et du plan de sondage dans lesquelles le logement est l'unité de sélection et la fourniture aux intervieweurs des numéros de téléphone afin de disposer d'une marge de manœuvre dans l'emploi d'une combinaison d'interviews téléphoniques et personnelles pour l'obtention des interviews du premier mois.

## BIBLIOGRAPHIE

CATLIN, G., CHOUDHRY, H., et HOFMANN, H. (1984). Telephone ownership in Canada. Document interne, Statistique Canada.

CATLIN, G., et INGRAM, S. (1988). The effects of CATI on cost and data quality. *Telephone Survey Methodology*, (éds. R. Groves et coll.), 437-450. New York: Wiley.

CHOUDHRY, G.H. (1984). Results from telephone interviewing experiment in the non self representing areas of the Labour Force Survey. Document interne, Statistique Canada.

DREW, J.D., et GAMBINO, J. (1991). Plans for the 1991 post censal redesign of the Canadian Labour Force Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, à paraître.

DREW, J.D., CHOUDHRY, H., et HUNTER, L. (1988). Nonresponse issues in government telephone surveys. *Telephone Survey Methodology*, (éds. R. Groves et coll.), 233-246. New York: Wiley.

DREW, J.D., DICK, P., et SWITZER, K. (1989). Development and testing of telephone survey methods for household surveys at Statistics Canada. *Proceedings of the Section on Survey Research Methods, American Statistical Association*.

DREW, J.D., et JAWORSKI, R. (1986). Telephone survey development on the Canadian Labour Force Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*.

DREW, J.D., ROYCE, D., et van BAAREN, A. (1989). Address register research at Statistics Canada. *Proceedings of the Section on Survey Research Methods, American Statistical Association*.

DUFOUR, J. (1990). Implantation de la première entrevue par téléphone pour la base d'appartements de l'enquête sur la population active. Document interne, Division des méthodes d'enquêtes sociales, Statistique Canada.

GROVES, R.M., BIEMER, P.P., LYBERG, L.E., MASSEY, J.T., NICHOLLS, W.L., II, et WAKSBERG, J. (éds) (1988). *Telephone Survey Methodology*. New York: Wiley.

opérations d'enquête et les saisies de données régionales. Par contre, les travaux de remaniement pendant les années 80 se sont limités à l'échantillon.

Même si les décisions à propos de l'ampleur du remaniement post-censitaire de 1991 n'ont pas encore été prises, tout programme qui se situerait entre la révision majeure des années 70 et le remaniement minimal des années 80 semble nécessaire. Les travaux du remaniement en sont encore au début de la planification, et ils passeront par quatre sous-projets portant sur: i) le contenu et le questionnaire, ii) la modernisation des systèmes de traitement de l'enquête et l'examen des produits de l'enquête, iii) l'élaboration, les tests et la mise en oeuvre des interviews assistées par ordinateur et iv) le remaniement de l'échantillon. (Drew et coll. 1991).

Les sous-projets iii) et iv) sont ceux qui portent sur les méthodes téléphoniques et IAO. Les plans actuels pour ces sous-projets dans la mesure où ils se rattachent aux facteurs d'étude de l'enquête exposés dans les sections précédentes de cette communication sont résumés brièvement ci-dessous.

### Technologie et organisation

Compte tenu des résultats positifs des essais des interviews assistées par ordinateur pour l'EPA mentionnés par Catlin et coll. (1988), on a décidé de faire de l'IAO un des principaux éléments du remaniement. De plus, pour des raisons déjà exposées, on a décidé que l'organisation locale actuelle des intervieweurs devrait être conservée, et la mise en oeuvre des méthodes IAO prendra la forme d'interviews personnelles assistées par ordinateur (IPAO). Plus précisément, les intervieweurs locaux recevront des ordinateurs portables légers qu'ils prendront avec eux pour les interviews personnelles et ils feront des interviews téléphoniques à partir de leur maison. La combinaison des interviews personnelles et téléphoniques pourrait rester en gros la même qu'actuellement, soit 80% par téléphone et 17% personnelles, mais on pourrait augmenter le nombre d'interviews téléphoniques si l'on décidait d'interviewer davantage par téléphone à froid les ménages pendant leur premier mois dans l'échantillon.

Le plan de travail du sous-projet des interviews assistées par ordinateur comprend un essai sur le terrain en 1991 d'un ordinateur portable à écran par touches, et au cours des années suivantes, l'élaboration ou l'acquisition du logiciel IAO, un essai combiné des autres méthodes IAO et de questionnaire, l'élaboration de l'édition en direct et une version automatisée du manuel de l'intervieweur incorporée dans un écran HELP accessible pendant l'interview.

### Base de sondage et mode de collecte

Une conclusion fondamentale des recherches a été que les interviews téléphoniques à froid avec un suivi personnel donnaient des taux de réponse et des estimations de la population active comparables à celles obtenues avec la procédure de collecte téléphonique à chaud actuelle pour l'EPA, qui comprend une interview personnelle pour les ménages du premier mois dans l'échantillon. On a observé une exception au Québec, où, comme on l'a déjà dit, des différences dans les données sont apparues pendant une période. L'importance du suivi personnel dans le maintien des taux de réponse favorise le maintien du logement comme unité de sondage et les numéros de téléphone fournis aux intervieweurs leur donneront plus de souplesse pour l'utilisation du téléphone et les interviews personnelles pour l'obtention des interviews du premier mois.

Dans les régions urbaines, la base aréolaire actuelle et le registre des adresses comme base de sondage sont cohérents avec cette approche. Dans les régions rurales, comme on l'a déjà dit, on étudiera la faisabilité d'apparier les adresses de la base de sondage aréolaire et les listes téléphoniques afin de fournir aux intervieweurs les numéros de téléphone. On prévoit également de poursuivre l'étude des méthodes de bases de sondage duales qui, dans les régions urbaines, pourraient comprendre le registre des adresses et une base de sondage aréolaire, et dans les régions rurales, une base de sondage aréolaire et une base de sondage téléphonique.



On envisage pendant le remaniement post-censitaire de 1991 de l'enquête sur la population active d'effectuer des études sur l'utilisation du registre des adresses comme base de sondage des enquêtes-ménages dans les régions urbaines. Si l'on en conclut qu'on devrait l'adopter comme base de sondage, le registre des adresses sera mis à jour sur une base permanente après le recensement de 1991.

Un avantage du registre des adresses comme base de sondage sur la base de sondage aréolaire est que les ménages avec et sans téléphone sont connus à l'avance. Les deux par conséquent peuvent être échantillonnés comme des strates distinctes, avec un nombre de grappes réduit pour la strate téléphonique, pour laquelle une partie importante des interviews du premier mois peut être faite par téléphone. La strate non téléphonique comprendrait les ménages avec les numéros non publiés et les ménages sans téléphone. D'après les études antérieures, on peut constater que le taux de refus des interviews téléphoniques à froid des ménages avec un numéro non publié est de 12%, comparativement à 4% pour tous les ménages. Dans le cas des interviews téléphoniques à chaud, il n'y a pas d'augmentation correspondante des taux de refus pour les ménages avec des numéros non publiés, et le taux de conversion de ces ménages pour répondre au téléphone au cours des mois suivants est bon. Cette constatation et le désir de respecter les inquiétudes concernant le respect de la vie privée viennent favoriser les interviews personnelles pour ces ménages, qui représentent le pourcentage estimatif de 10-15% des numéros.

### Base de sondage duale

Dans les régions urbaines, si la couverture du registre des adresses comme base de sondage unique ne convient pas, on peut envisager un plan de sondage dual dans lequel le registre des adresses est complété par un échantillon régional. Un tel échantillon complémentaire pourrait prendre plusieurs formes. Une option intéressante, qui permet de se passer des dépenses de la construction et de la tenue d'une base de sondage aréolaire ordinaire et d'un registre des adresses, serait d'utiliser une approche par intervalle dans laquelle un échantillon de logements consécutifs dans le registre des adresses serait choisi et vérifié sur le terrain. Tout logement manquant dans le registre des adresses

Milan (1990) a étudié les méthodes de bases de sondage duales en considérant une optimisation des coûts et de la variance pour le cas général dans lequel aucune des bases ne doit recouvrir l'univers complet. On a estimé que ceci serait un modèle pratique puisque la base de sondage aréolaire, même si elle est complète sur le plan conceptuel, dans la pratique se caractérise par un sous-dénombrement de 3-4% par rapport au recensement, en plus du 5% de la population qui n'est pas représentée en raison de la non-réponse. L'extension du modèle de Milan afin d'inclure une composante d'erreurs non due à l'échantillonnage permettrait de factoriser dans l'optimisation ce que nous savons ou ce que nous désirons supposer à propos des biais de couverture et de non-réponse avec d'autres bases de sondage. Il peut être utilisé dans le contexte des bases de sondage duales qui combinent le registre des adresses et les bases de sondage aréolaires dans les régions urbaines et la combinaison des bases de sondage aréolaires et téléphoniques dans les régions rurales.

## 7. REMANIEMENT POST-CENSAIRE DE 1991 DE L'EPA

L'enquête sur la population active est remaniée après chaque recensement décennal de la population. Le remaniement porte habituellement sur le remaniement de l'échantillon, mais dans les années 70, on a effectué une révision importante qui comprenait le remaniement de l'échantillon, des modifications au questionnaire, au libellé des questions et aux résultats de l'enquête et une refonte poussée des systèmes de traitement de l'enquête, y compris la mise en place d'un réseau de mini-ordinateurs dans les bureaux régionaux afin d'appuyer les



Un autre inconvénient des bases de sondage téléphoniques, en particulier pour les enquêtes par panel, est qu'elles se détériorent rapidement. Drew, Dick et Switzer (1989) ont trouvé un taux d'addition et de suppression de 0,5 – 1,0% au stock des numéros résidentiels publiés par mois. Les échantillons téléphoniques ne peuvent donc rester représentatifs de l'univers téléphonique pendant très longtemps à moins d'être mis à jour. Les auteurs proposent une stratégie de mise à jour des échantillons d'une enquête par panel en utilisant pour cela des fichiers de numéros publiés acquis sur une base permanente auprès des compagnies de téléphone. Un test opérationnel de la procédure sur une période de neuf mois s'est révélé concluant. Leur procédure ne porte que sur les numéros publiés échantillonnés seulement provenant d'une base de sondage par listes et ne constitue pas une solution au problème de la tenure à jour d'un échantillon utilisant les méthodes de composition téléphonique aléatoire pendant la durée de vie d'un panel.

En raison des problèmes de couverture et de mise à jour des bases de sondage téléphoniques, on estime que les approches utilisant les logements plutôt que les numéros de téléphone comme unités d'échantillonnage sont plus prometteuses pour les grandes enquêtes par panel. Il est intéressant de noter que la situation peut être assez différente pour d'autres enquêtes. Catlin et coll. (1984) ont montré que les biais de couverture des caractéristiques de la population générale étaient moindres que pour les caractéristiques de la population active. De plus, pour les enquêtes plus petites (taille de l'échantillon de 10,000 ou moins), les petits biais sont moins importants compte tenu de l'erreur d'échantillonnage relative plus grande pour ces enquêtes. On a ainsi été amené à établir une enquête auprès des ménages CTA en 1986. Elle a servi à de nombreuses enquêtes uniques et pour l'enquête sociale générale, qui est une enquête annuelle auprès de 10,000 ménages.

### Base de sondage aréolaire

Comme on l'a déjà dit, il est possible dans les régions urbaines d'appartier certaines adresses tirées d'une base de sondage aréolaire aux listes téléphoniques afin de procéder à des interviews téléphoniques à froid. L'expérience avec l'EPA a permis de constater que l'on pouvait obtenir des numéros de téléphone de cette façon pour environ 60% des ménages. Ces taux d'appariement reposent sur un appariement exact après redressement des renseignements de l'adresse et il est possible de les améliorer grâce à des méthodes de raccordement des enregistrements. En téléphonant à une partie appréciable des cas du premier mois, la mise en grappes de l'échantillon pourrait être réduite quelque peu, mais un échantillon en grappes reste une contrainte de tout plan de sondage aréolaire.

On envisage d'étudier la possibilité d'étendre ces procédures aux régions rurales, ce qui permettrait de changer le type de renseignements recueillis lorsqu'on crée des listes de logements pour l'enquête. Les renseignements actuellement recueillis ont tendance à être une description des caractéristiques physiques des logements, alors que si l'on désire effectuer un appariement avec les listes des abonnés au téléphone, il faudrait avoir des renseignements tels que le nom (qui se trouve souvent sur les boîtes à lettres), le nom et le numéro de la rue, ou en l'absence de ces renseignements, le numéro de la route rurale et le code postal.

### Registre des adresses

Statistique Canada est en train de construire un registre des adresses dans les régions urbaines du Canada. Il servira pendant le recensement de 1991 à améliorer la couverture en fournissant une vérification indépendante des listes de logements créées par les recenseurs (Drew, Royce et van Baaren 1989). Le registre des adresses sera une liste lisible par machine des adresses obtenue par le raccordement de diverses sources de données administratives, y compris les listes de clients avec les numéros publiés achetés auprès des compagnies de téléphone. Pendant l'utilisation du registre des adresses au cours du recensement de 1991, on mettra à jour sa couverture afin de la faire correspondre à celle du recensement de 1991.

Tableau 4  
Ménages sans téléphone par province (%)

	1976	1981	1985	1987	1990
Canada	3.5	2.4	1.8	1.5	1.5
Terre-Neuve	10.0	6.0	5.1	3.6	1.9
Ile-du-Prince-Edouard	-	-	-	-	2.8
Nouvelle-Ecosse	7.5	4.6	3.5	3.2	1.5
Nouveau-Brunswick	5.8	5.3	5.3	3.3	2.2
Québec	3.3	2.1	1.6	1.5	1.5
Ontario	2.5	1.9	1.0	1.0	1.2
Manitoba	4.1	2.3	2.7	2.4	1.7
Saskatchewan	3.6	2.5	2.3	2.4	2.3
Alberta	3.0	2.4	2.0	1.8	2.0
Colombie-Britannique	4.2	2.8	2.4	1.3	1.5

Source: Statistique Canada, estimations provenant de l'enquête sur les installations et l'équipement ménagers.

Tableau 5

Caractéristiques de la population active selon le statut téléphonique

Province	Statut téléphonique	Taux de chômage	Taux d'activité
Nouvelle-Ecosse	Publié	9.0	71.9
	Non publiés	9.8	70.2
	Pas de téléphone	17.2	62.3
	Publiés	6.3	80.7
Alberta	Publiés	8.2	81.5
	Non publiés	11.1	67.0
	Pas de téléphone		
	Publiés		

Le tableau 4 donne le pourcentage des ménages sans téléphone au Canada de 1976 à nos jours. La couverture téléphonique, bien que déjà élevée en 1976, a continué de s'étendre, même s'il semble qu'elle se soit stabilisée au cours des dernières années à environ 98,5%.

Laflamme (1990) a entrepris une étude comparative des caractéristiques des univers sans téléphone et avec téléphone. L'enquête comprenait une ventilation de ceux ayant un numéro publié et de ceux n'ayant pas de numéro publié, obtenus par le raccordement des numéros de téléphone fournis par les répondants EPA à des listes de numéros téléphoniques publiés. Deux provinces ont été incluses dans l'étude, la Nouvelle-Ecosse et l'Alberta. On devait constater que pour la Nouvelle-Ecosse et l'Alberta, 9,7% et 11,9% respectivement des numéros n'étaient pas publiés. Les taux de chômage et d'activité calculés par Laflamme figurent au tableau 5. Cette étude reproduit les conclusions d'études antérieures selon lesquelles les caractéristiques de la population active des personnes sans téléphone sont très différentes de celles qui en ont un. Les caractéristiques de la population active diffèrent, mais à un degré moindre, entre les personnes qui ont des numéros publiés et celles qui ont des numéros non publiés.

Bien que la population sans téléphone représente 1,0-1,5% de la population seulement, les différences dans les caractéristiques de la population active sont suffisamment importantes pour que la simple exclusion de la population sans téléphone ne soit pas une option valable pour l'EPA compte tenu de la précision nécessaire pour les estimations nationales de l'emploi et du chômage (coefficients de variation de 0,5% et de 2% respectivement).



coûts de collecte des données, donne une augmentation de 68-75% de la non-réponse par rapport à l'organisation locale. Comme on l'a vu en 3), il y a des indices que cette non-réponse supplémentaire introduit un biais de non-réponse sérieux dans les estimations EPA. De plus, on s'inquiète de ce que l'écart des taux de non-réponse entre les organisations locales et centrales pourrait s'élargir à l'avenir, puisque une sollicitation téléphonique plus forte et la disponibilité accrue des techniques de tri téléphonique rendent la population moins réceptive aux interviews téléphoniques. Pour ces raisons, on favorise des stratégies de plan de sondage qui, même si elles sont souples pour ce qui est des appels téléphoniques, peuvent permettre aussi des suivis personnels au besoin. L'organisation locale est celle qui offre la flexibilité la plus grande. Pour toutes ces considérations, on a décidé de conserver l'organisation locale actuelle.

## 5. TECHNOLOGIE

Catlin, Ingram et Hunter (1988) ont effectué une étude contrôlée comparant les méthodes ITAO et celle du papier et du crayon. Dans l'étude, on a soumis le questionnaire EPA à des échantillons CTA de 1,000 ménages par mois par traitement sur une période de neuf mois. Toutes les interviews ont eu lieu à partir du siège central de Statistique Canada à Ottawa. L'étude faisait partie d'un travail de recherche en collaboration avec le United States Bureau of the Census (USBC) et le logiciel ITAO utilisé a été mis au point par ce dernier. On a délibérément retenu le même libellé des questionnaires pour les deux approches. Les caractéristiques propres à l'approche ITAO sont le branchement automatique, quelques vérifications en direct sur deux bases et l'acheminement automatique des appels.

On a discerné trois améliorations qualitatives pour l'ITAO par rapport à l'autre méthode. D'abord, le taux général de rejet à la vérification pendant le traitement post-collecte des données était inférieur de 50% pour l'ITAO. Ensuite, il y a eu une disparition pratiquement totale des erreurs de branchement. Ce qui est important, c'est que cette baisse s'observe pour certaines parties du questionnaire, qui, même si elles sont rarement utilisées, ont une incidence sur la détermination de la situation vis-à-vis de l'activité, et pour laquelle des intervieweurs utilisant le papier et le crayon font des erreurs de branchement importantes. Enfin, la taille du ménage moyen déclarée pour l'ITAO était supérieure de 3%, ce qui représente en gros une réduction de 50% du sous-dénombrement de l'EPA par rapport au recensement. Cette amélioration semble provenir du mécanisme de contrôle incorporé dans l'instrument ITAO pour les membres supplémentaires du ménage et pour les personnes temporairement absentes.

Compte tenu de ces résultats, on a décidé que la mise en oeuvre des interviews assistées par ordinateur serait une des principales caractéristiques du remaniement post-censitaire de 1991 de l'EPA. En raison de la préférence pour le maintien d'une organisation d'intervieweurs locale, on prévoit une mise en oeuvre IPAO.

## 6. BASE DE SONDAGE ET PLAN DE SONDAGE

### Base de sondage téléphonique

La couverture téléphonique et la mesure dans laquelle des caractéristiques des personnes sans téléphone diffèrent de celles de ceux qui en ont un sont des facteurs importants dans l'étude de méthodes d'enquêtes téléphoniques, en particulier pour ce qui concerne les stratégies de la base de sondage.

Dans une revue internationale de la couverture téléphonique, Trewin et Lee (1988) ont constaté que la couverture téléphonique au Canada est l'une des plus élevées du monde, avec 97-98%. Comme c'est la règle générale dans la plupart des pays qu'ils ont étudiés, les personnes vivant dans des ménages sans téléphone au Canada ont tendance à avoir des revenus plus bas et des taux de chômage plus élevés.



actuel des interviews téléphoniques à chaud était conservé. L'organisation mixte a fait l'objet d'un essai entre janvier 1988 et mars 1989 dans deux régions métropolitaines de recensement où se trouvaient des bureaux régionaux, Montréal et Halifax. Son premier objectif était de mesurer l'élément coût d'une telle organisation.

La méthodologie des tests consistait en une collecte personnelle par les intervieweurs locaux pour les ménages du premier mois dans l'échantillon et par des interviews téléphoniques par le personnel central travaillant dans les bureaux régionaux pour la plupart des autres cas. Lorsqu'un suivi de non-réponse était nécessaire pour les ménages affectés au départ aux intervieweurs centraux, celui-ci était effectué par les intervieweurs locaux. Cette méthodologie avait été testée au départ pour l'enquête sur la population active par Muirhead et coll. (1975) et a fait l'objet d'études poussées par l'United States Bureau of the Census (1987), où les interviews centrales se font grâce à des interviews téléphoniques assistées par ordinateur (ITAO). Une des complexités de la méthode était la pratique reconnue pour la première moitié du test de transférer des cas nécessitant un suivi de non-réponse des intervieweurs centraux aux intervieweurs locaux au milieu de la semaine d'interview. Pour la deuxième moitié du test, ce recyclage était limité aux cas où le numéro de téléphone n'était plus valide. Pendant la première moitié du test, les taux de non-réponse étaient de 8,0% pour le traitement de test contre 6,1% pour les procédures de contrôle correspondant à l'interviews décentralisée utilisée pour l'EPA courante. L'écart s'est réduit à 7,3% contre 6,7% pendant la deuxième moitié.

À partir du premier test d'échantillonnage téléphonique, on a estimé que les coûts de l'interview par ménage étaient de \$2,72 pour la collecte centrale de données avec un échantillonnage de listes téléphoniques, contre \$3,53 pour l'échantillonnage CTA. Les coûts supplémentaires des méthodes CTA s'expliquent par le temps consacré au tri des numéros de téléphone résidentiels. Ces coûts comprennent \$0,46 par ménage pour les appels interurbains. On a estimé ce montant à partir des taux interurbains et des données sur la durée des appels, puisque les pratiques de tenue des livres dans les bureaux régionaux ne permettent pas l'obtention des coûts réels. Les coûts comparables pour l'EPA courante sont de \$4,76 par ménage pour les frais d'intervieweurs et les dépenses. Le test de l'organisation mixte a donné une économie par rapport à l'EPA courante de \$0,78 par ménage en frais d'intervieweurs et dépenses. La comparaison des coûts ci-dessus ne prend pas en compte le coût des locaux à bureaux et du matériel dans les coûts des organisations centralisées et mixtes. Elle ne prend également pas en compte le coût du transfert des documents en direction et en provenance des intervieweurs locaux dans les organisations mixtes et locales, ce qui, dans la technologie actuelle du papier et du crayon, se fait par l'envoi par messagerie des documents, mais se ferait de façon électronique dans les options CTA.

On a envisagé l'organisation mixte pour les villes avec bureau régional seulement, puisque l'extension de l'une de ces dernières se traduirait par une utilisation accrue des appels téléphoniques interurbains. Qui plus est, le plan de sondage des régions urbaines et rurales plus petites est en grappes, de sorte que les unités primaires d'échantillonnage donnent des tailles d'échantillon qui correspondent à une tâche d'intervieweur. La centralisation de la partie téléphonique de l'échantillon nécessiterait une mise en grappes plus forte de l'échantillon afin de conserver une charge de travail suffisante pour les intervieweurs locaux. Par ailleurs, dans les centres urbains de taille moyenne, où il y a actuellement de quatre à cinq intervieweurs, le nombre d'intervieweurs locaux dans une organisation mixte serait ramené à un ou deux, ce qui réduirait sensiblement la marge de manœuvre quand les intervieweurs devront remplacer les autres pendant les vacances et les maladies.

Tous les trois modèles organisationnels considérés ont des avantages et des inconvénients. L'organisation locale a les taux de non-réponse les plus bas, mais le coût de collecte unitaire des données le plus élevé. L'organisation mixte a une non-réponse un peu plus élevée et des coûts un peu plus bas, et se trouve limitée dans son champ d'application. En fin de compte, on a exprimé que l'organisation mixte entraînerait bien des complexités pour un gain tout au plus minime. L'organisation centrale, qui se caractérise par des économies appréciables des

Tableau 3  
Test de l'échantillonnage téléphonique (Octobre 1985 - septembre 1986)  
Taux de chômage et d'activité

Province		Plan		Taux de chômage (D.S.)		Taux d'activité (D.S.)	
Québec	LISTE	12.3	(0.78)	64.1	(1.08)	62.8	(1.28)
	CTA	13.0*	(0.88)	63.4	(0.29)	69.0	(1.11)
	EPA	10.9	(0.27)	69.0	(1.18)	69.0	(0.20)
	LISTE	7.3	(0.59)	69.0	(1.11)	69.0	(1.18)
Ontario	LISTE	7.3	(0.59)	69.0	(1.11)	69.0	(1.18)
	CTAI	7.9	(0.63)	69.0	(1.18)	69.0	(1.18)
	EPA	6.9	(0.16)	69.0	(1.18)	69.0	(1.18)
	LISTE	6.9	(0.16)	69.0	(1.18)	69.0	(1.18)

\* Différence significative entre les taux de chômage CTA et EPA pour le Québec.

des taux de chômage n'étaient pas statistiquement significatives, les taux pour les interviews téléphoniques à froid étaient plus élevés. D'autres chercheurs ont observé des différences dans le même sens, mais également n'ont pu être en mesure d'attribuer une signification statistique à celles-ci. Ces données pourraient bénéficier d'une méta-analyse.

En résumé, les résultats du test révèlent que les interviews téléphoniques à froid sans suivi personnel donnent des taux de non-réponse plus élevés que la méthode actuelle des interviews téléphoniques à chaud, et même si cela n'est pas concluant, il y a des indices qu'il donne des taux de chômage plus élevés. Par contre, les interviews téléphoniques à froid avec suivi personnel, à l'exception de la seule période pour le Québec, ont donné des résultats comparables à ceux des interviews téléphoniques à chaud.

Compte tenu de ces résultats, on a décidé de mettre en oeuvre les interviews téléphoniques à froid avec suivi personnel pour l'échantillon d'appartements EPA qui représentent approximativement 4% de l'échantillon total. On a estimé que l'existence des numéros de téléphone des unités de la base de sondage des appartements aiderait à surmonter les problèmes de l'accès aux immeubles à étages multiples et permettrait de procéder à plus de tentatives de trouver les personnes à la maison que pour les interviews personnelles. Il semble que ces attentes se soient réalisées. Comme le signale Dufour (1990), alors que les taux de non-réponse du premier mois pour l'échantillon des appartements continuent de dépasser ceux du premier mois pour l'échantillon hors appartements, l'écart s'est réduit, passant d'une différence de 8.7 points au cours de l'année avant le changement à une différence de 5.7 points au cours des cinq premiers mois d'utilisation de la nouvelle procédure.

Un autre changement au mode de collecte pour l'EPA courante a été le suivi téléphonique des ménages du premier mois dans l'échantillon qui ne pouvaient contacter au cours d'une visite initiale. Cette procédure a été instaurée en 1986 et s'est traduite par une économie de \$100,000 dans les coûts de collecte des données.

L'effet combiné de la première interview téléphonique pour les appartements et du suivi téléphonique des non-répondants du premier mois a été une hausse du taux des interviews téléphoniques globales de l'enquête, de 80% en 1985 à 83% en 1990.

#### 4. ORGANISATION DES INTERVIEWS

Pendant le programme des tests, on a étudié deux solutions de rechange à l'organisation locale actuelle des interviews. Les tests des échantillons téléphoniques déjà décrits envisageaient une organisation "centrale" dans laquelle les bureaux régionaux faisaient toutes les interviews. Un autre test portait sur une organisation mixte, dans laquelle le mode de collecte



Tableau 2

Taux de non-réponse: interviews téléphoniques à chaud et interviews téléphoniques à froid avec/sans suivi personnel

Méthode	Test 1	Test 2
Interviews téléphoniques à chaud avec lettre (EPA courante)	4.1	5.6
Interviews téléphoniques à froid avec lettre	6.9	9.8
Interviews téléphoniques sans lettre	8.5	-

Test 1: Octobre 1985 – septembre 1986, Ontario et Québec.  
Test 2: Juillet 1988 – mars 1989, Nouvelle-Ecosse et Alberta.

exogènes à l'enquête se sont traduites par un contexte dans lequel les interviews téléphoniques à froid ont paru plus suspectes. Nous avons été heureux d'avoir pu effectuer le test pendant cette période, parce que la conclusion que les résultats de l'enquête obtenus par des interviews téléphoniques à froid sont davantage sujets à des influences extérieures que celles obtenues par des interviews téléphoniques à chaud sera une considération importante dans toute décision concernant l'extension des interviews téléphoniques.

On a également étudié les interviews téléphoniques à froid sans suivi personnel. Deux tests d'échantillonnage par téléphone ont été effectués, au cours desquels l'EPA a été exécutée comme une enquête téléphonique centrale avec des interviews à partir des bureaux régionaux. Le tableau 2 présente les taux de non-réponse pour les deux tests et les taux comparables pour l'EPA courante.

Le premier test a étudié deux méthodes d'échantillonnage: la CTA et une combinaison d'échantillonnage de liste pour les numéros publiés et la CTA pour les numéros non publiés. L'échantillonnage par liste contenait des lettres d'introduction, mais l'échantillonnage CTA non. Les différences des taux de réponse semblent mettre en évidence les effets positifs sur les taux de réponse d'une lettre préalable. Pour les deux tests, la comparaison des deux types d'interviews téléphoniques a révélé que les taux de non-réponse pour les interviews téléphoniques à froid étaient plus élevés, au niveau de signification de 5%. Le deuxième test a utilisé uniquement un échantillon de listes de numéros publiés.

Un problème important est le biais de non-réponse éventuelle, imputable à la non-réponse plus élevée caractéristique des interviews téléphoniques à froid sans suivi personnel. Comme approximation de ces non-répondants supplémentaires, Laflamme (1990) a considéré les ménages autres que ceux du premier mois de l'échantillon de l'EPA courante qui avaient un téléphone mais qui étaient interviewés de façon personnelle. Il avait constaté que la taille du groupe choisi, c'est-à-dire 3.5% des répondants, était proche de celle des cas de non-réponse supplémentaires pour les interviews téléphoniques à froid sans suivi personnel. De plus, il avait constaté que le taux de chômage du groupe de remplacement était de 12.8%, comparativement à 7.4% pour les personnes dans les ménages interviewés par téléphone. L'exclusion du groupe de remplacement de l'échantillon aurait réduit le taux de chômage national de 8.1% à 7.9%. Ceci constitue visiblement un biais sérieux, compte tenu de la précision nécessaire pour les estimations nationales de l'EPA. Comme l'hypothèse de remplacement semble raisonnable, ces conclusions soulèvent des préoccupations graves à propos des interviews téléphoniques à froid sans suivi personnel pour l'EPA.

Le tableau 3 compare les taux de chômage et d'activité pour le premier test d'échantillonnage téléphonique aux estimations correspondantes des ménages téléphoniques de l'EPA. La seule estimation qui devait se révéler significativement différente au niveau de 5% des estimations produites pour la population téléphonique de l'EPA courante était le taux de chômage pour le Québec pour le traitement CTA. Il est intéressant de noter que le test a été effectué en même temps que des problèmes apparaissaient pour les estimations du Québec dans le test de la première interview téléphonique. Un autre point intéressant est que si les autres différences



Tableau 1  
Test de la première interview téléphonique (octobre 1985 – mars 1989)  
Estimation du traitement du test en pourcentage du traitement de l'estimation de contrôle

Caractéristique	Ontario		Québec	
	Pourcentage	Statistique t	Pourcentage	Statistique t
Emploi	98.5	-1.22	97.2	-1.64
Chômage	96.3	-0.94	111.8	2.27*
Inactifs	101.1	0.88	98.2	-0.63
Pop. 15 +	909.2	-0.86	98.5	-1.31
Pop. ménages = 3 +	97.8	-0.76	94.1	-1.26
Pop. ménages 1 personne	100.6	0.25	104.1	-0.93
Pop. ménages 2 personnes	101.1	0.26	101.0	0.43
Pop. ménages 3 personnes	98.6	-0.56	94.7	-1.28
Hommes occupés 15-24	95.2	-0.75	88.7	-2.00*
Hommes occupés 25 +	98.5	-1.11	96.7	-1.54
Femmes occupées 15-24	99.0	-0.11	111.5	1.44
Femmes occupées 25 +	99.2	-0.65	97.1	-1.11
Hommes chômeurs 15-24	105.6	0.53	119.0	1.53
Hommes chômeurs 25 +	94.5	-0.80	96.7	-0.31
Femmes chômeurs 15-24	99.6	-0.15	119.4	1.30
Femmes chômeurs 25 +	90.9	-1.22	123.9	2.71**
Inactifs 15-24	99.8	-0.07	99.9	0.47
Inactifs 25 +	105.6	1.57	101.3	0.07
Inactives 15-24	101.9	0.56	95.3	-0.56
Inactives 25 +	99.2	-0.14	97.3	-0.92
Hommes 15-24	97.2	-0.49	95.2	-0.73
Hommes 25 +	99.9	-0.09	97.7	-1.49
Femmes 15-24	99.9	0.16	105.4	0.95
Femmes 25 +	98.9	-1.28	98.4	-1.24

\* Statistique t significative au niveau de 5%.  
\*\* Statistique t significative au niveau de 1%.

échantillon de contrôle et ont suivi les procédures habituelles pour tous les logements dont le numéro de téléphone n'avait pas été communiqué.

On n'a pas trouvé de différence significative des taux de réponse entre les échantillons de test et de contrôle. Pour le Québec, les taux de réponse étaient de 96,1% pour les deux échan-

tilions, tandis que pour l'Ontario, le taux de 96,3% pour l'échantillon test était très légèrement inférieur à celui de 96,5% pour l'échantillon de contrôle.

La comparaison des estimations de la population active obtenues dans les échantillons de test et de contrôle a permis de constater que certaines estimations du Québec au début du test pour la période octobre 1985 – février 1987 étaient significativement différentes. En particulier, le nombre d'hommes occupés et de chômeurs dans les ménages de trois personnes ou plus était sous-estimé dans l'échantillon de test (Drew, Choudhry et Hunter 1988). Le tableau 1 présente les données pour l'ensemble de la période des tests (octobre 1985 – mars 1989). Pour le Québec, quelques différences statistiquement significatives existent, résultant de l'influence de la période antérieure. Lorsqu'on a analysé les données à compter de mars 1987, ces différences ne sont pas significatives. En Ontario, on n'a relevé aucune différence significative. S'agissant des différences pour le Québec, leur coïncidence avec un programme d'inspection des bénéficiaires du Bien-être social effectué par l'administration provinciale semble indiquer que des mesures

En 1985, après la mise en oeuvre d'un échantillon remanié, on a entrepris un programme de recherche afin d'étudier quelles améliorations pour la collecte des données pourraient résulter i) d'un accroissement de l'utilisation du téléphone pour la collecte, ii) des méthodes d'enquête par téléphone où le téléphone sert à l'échantillonnage et à la collecte des données, et iii) des méthodes d'interview assistée par ordinateur (IAO).

Dans l'étude des méthodes téléphoniques et IAO, on a constaté qu'il était utile de caractériser le plan de sondage en termes d'un certain nombre de facteurs comme suit:

i) **Mode de collecte.** Comme on l'a déjà dit, le mode de collecte actuel est l'enquête téléphonique à chaud, avec une interview personnelle initiale et des interviews essentiellement téléphoniques au cours des mois suivants. D'autres modes de collecte comprennent les interviews téléphoniques à froid avec un suivi personnel des non-répondants téléphoniques et les interviews téléphoniques à froid sans suivi personnel.

ii) **Organisation des interviews.** L'EPA a actuellement une organisation locale, et les intervieweurs locaux font un mélange d'interviews personnels et d'interviews téléphoniques à partir de leur maison. Une autre possibilité est une organisation centrale, les intervieweurs travaillant à partir d'une ou de plusieurs villes centrales; dans le cas de Statistique Canada, il s'agit de ses cinq bureaux régionaux au pays. Un troisième modèle organisationnel est un modèle combiné, dans lequel l'interview se fait par une combinaison d'intervieweurs locaux et centraux.

iii) **Technologie.** La technologie actuelle de l'enquête est le papier et le crayon. L'autre technologie que l'on considère plus valable pour les enquêtes-ménages est l'interview assistée par ordinateur. L'IAO est habituellement appelée ITAO (interview téléphonique assistée par ordinateur) lorsqu'elle est faite de façon centrale et IPAO (interview personnelle assistée par ordinateur) lorsqu'elle est faite localement en utilisant des ordinateurs portables que les intervieweurs utilisent pour les interviews personnelles et pour les interviews téléphoniques à partir de leur maison.

iv) **Base de sondage et plan de sondage.** L'enquête utilise un plan de sondage aréolaire depuis son début, en 1945. Les autres possibilités comprennent les bases de sondage téléphoniques utilisant les méthodes de composition aléatoire (Waksberg 1978) ou une combinaison de bases de sondage, contrairement aux listes téléphoniques, sont conceptuellement complètes, et les méthodes de sondage duales combinant deux ou plusieurs des options ci-dessus.

Dans les sections qui suivent, on examine les résultats du programme de recherche et d'essais pour chacun de ces facteurs.

### 3. MODE DE COLLECTE

On a procédé à un important essai de 1985 à 1989 pour déterminer l'incidence des interviews téléphoniques à froid avec suivi personnel comme solution de remplacement aux interviews téléphoniques à chaud actuelles. Cet essai s'appelle l'essai téléphonique de la première interview. Il a été incorporé dans l'EPA courante dans les régions urbaines du Québec et de l'Ontario. Drew, Choudhry et Hunter (1988) en ont exposé la méthodologie détaillée. En bref, les logements EPA nouvellement échantillonnés sont apparés aux listes achetées des compagnies téléphoniques à partir de l'adresse. On a obtenu des taux d'appariement de 65%. On a choisi des échantillons d'essai et de contrôle à partir des logements apparés de la même unité d'échantillement d'essai à été couplé à un logement de contrôle à l'intérieur de la même unité d'échantillement (lot urbain). Pour les logements tests, on a donné les numéros de téléphone aux intervieweurs qui devaient essayer de mener une interview téléphonique, mais faire appel à un suivi personnel au besoin. Les intervieweurs n'étaient pas au courant de l'existence d'un

# Recherche et essais pour les méthodes d'enquêtes par téléphone à Statistique Canada

J. DOUGLAS DREW<sup>1</sup>

## RÉSUMÉ

On présente les résultats de la recherche et des essais des méthodes d'enquêtes par téléphone et assistées par ordinateur pour les enquêtes-ménages, que suit une discussion de la façon dont ces conclusions vont influencer le remaniement des enquêtes-ménages à Statistique Canada pendant les années 90. On s'attache plus particulièrement dans la population active du Canada.

MOTS CLÉS: Collecte des données; enquêtes-ménages; plan de sondage.

## 1. INTRODUCTION

Les enquêtes ont subi au cours des années 80 d'importants changements en raison des progrès de la technologie et de l'élaboration de méthodes d'enquête par téléphone modernes et il est probable que le rythme de ces changements va s'accélérer pendant les années 90. Dans cette communication, nous décrivons les recherches, les essais et l'élaboration des méthodes qui constitueront l'infrastructure des activités de collecte des données pour les enquêtes-ménages de Statistique Canada pendant les années 90. Cette recherche s'est attachée en particulier à l'enquête sur la population active du Canada et a été menée de 1985 à 1989 afin d'identifier les améliorations à mettre en oeuvre pendant le remaniement post-censitaire de 1991 de l'enquête.

## 2. PROGRAMME DE RECHERCHE ET D'ESSAIS DE L'EPA

L'enquête sur la population active (EPA) du Canada est l'enquête-ménage la plus importante de Statistique Canada, avec un échantillon de 62,300 ménages chaque mois. Elle utilise un panel rotatif dans lequel les ménages restent dans l'échantillon pendant six mois consécutifs, avant d'être retirés. Elle utilise un sondage aréolaire multiple, avec un personnel décentralisé de 1,000 intervieweurs locaux situés au Canada et faisant rapport à un des cinq bureaux régionaux. Jusqu'au début des années 70, toutes les interviews se faisaient en personne. En 1972, on a mis en oeuvre les interviews téléphoniques dans les grandes régions urbaines pour les interviews de suivi auprès des ménages, après que ces derniers aient passé l'interview personnelle pendant leur premier mois dans l'échantillon. On appelle ce suivi téléphonique dans la littérature "enquête téléphonique à chaud" pour la distinguer des "enquêtes téléphoniques à froid", lorsque l'interview téléphonique n'est pas précédée d'une interview personnelle (Groves et coll. 1988). L'enquête téléphonique à chaud était au début limitée aux principales régions urbaines en raison du nombre de lignes groupées dans les régions urbaines et rurales plus petites et des préoccupations entourant la confidentialité des données recueillies. Toutefois, pendant le remaniement de l'enquête de 1981, on a essayé des interviews téléphoniques à chaud pour les petites régions urbaines et rurales, mais on a constaté que les répondants étaient prêts à être interviewés par téléphone et que la procédure n'avait pas d'incidence sur les taux de réponse ou les estimations d'enquêtes (Choudhry 1984). L'extension de l'interview téléphonique à ces régions en 1984 s'est traduite par une baisse de 10% des coûts de collecte des données de l'enquête.

<sup>1</sup> J. Douglas Drew, directeur adjoint, Division des enquêtes-ménages, Statistique Canada, Ottawa, Ontario, K1A 0T6.





Nous n'avons pas parlé de la possibilité d'exporter des données et des métadonnées du système Blaise vers le progiciel de base de données Paradox. Nous projetons aussi d'établir un lien entre Blaise et le système de gestion de base de données Oracle. Ainsi, il sera possible de créer une architecture client-serveur pour les utilisateurs de Blaise.

À l'extrémité de la chaîne de production des données statistiques, il y a encore quelques aspects de la publication qui méritent notre attention. En premier lieu, il faudra élaborer un logiciel qui permettra d'évaluer les risques de divulgation d'information confidentielle qui accompagnent la publication de données statistiques. On offrira aussi des "fonctions" qui permettront de mettre les tableaux ou les fichiers à l'abri de ces risques. Enfin, les organismes statistiques se livrent de plus en plus à la publication électronique de données statistiques (par ex.: information sur disquette, sur disque compact-ROM, etc.). Pour aider les utilisateurs de ce genre de support à choisir les séries de données dont ils ont besoin, il est nécessaire de mettre à leur disposition des logiciels conviviaux. C'est ce à quoi nous nous occupons présentement.

## BIBLIOGRAPHIE

- BETHLEHEM, J.G., VAN BUITENEN, A.A.A., HUNDEPOOL, A.J., ROESINGH, M.J., et VAN DE WETERING, A. (1989a). Abacus 1.0, A Tabulation Package, Compact Guide. CBS report, Voorburg: Netherlands Central Bureau of Statistics.
- BETHLEHEM, J.G., HUNDEPOOL, A.J., SCHUERHOFF, M.H., et VERMEULEN, L.F.M. (1989b). Blaise 2.0/An Introduction. CBS report, Voorburg: Netherlands Central Bureau of Statistics.
- BETHLEHEM, J.G., HUNDEPOOL, A.J., SCHUERHOFF, M.H., et VERMEULEN, L.F.M. (1989c). Blaise 2.0/Language Reference Manual. CBS report, Voorburg: Netherlands Central Bureau of Statistics.
- BETHLEHEM, J.G., et KELLER, W.J. (1987). Linear Weighting of Sample Survey Data. *Journal of Official Statistics* 3, 141-154.
- KELLER, W.J., BETHLEHEM, J.G., et METZ, K.J. (1990). The impact of microcomputers on survey processing at the Netherlands Central Bureau of Statistics. *Proceedings of 1990 Annual Research Conference, U.S. Bureau of the Census*, 637-645.

données d'enquête, nous devons trouver des "Fonctions" qui nous permettent d'exporter des données du système Blaise vers ces progiciels. Cela se fait en deux étapes. Premièrement, le fichier de données, en format Blaise, est converti en format ASCII et deuxièmement, l'information relative aux variables, qui est consignée dans le questionnaire Blaise, est traduite de manière à pouvoir être comprise par le progiciel en question. Un fichier provisoire est donc créé. En exécutant ce fichier à partir du progiciel statistique, on se trouve à créer un fichier informatique. Et en chargeant ce dernier, l'utilisateur peut débuter immédiatement son analyse sans se préoccuper de définir les variables, les étiquettes, etc.

La procédure que nous venons de décrire ne peut être réalisée qu'avec les progiciels SPSS et Stata et aucun autre. Nous pourrions, bien sûr, élaborer la même procédure pour chaque progiciel statistique connu mais cela exigerait un long travail de programmation. Nous avons choisi une autre solution. Le système Blaise contient un utilitaire spécial, générateur de fichiers provisoires. L'utilisateur "colore" la structure du fichier provisoire dans un logiciel de traitement de texte et exécute le générateur de fichiers provisoires en utilisant comme données d'entrée la description générale qu'il vient de faire du fichier provisoire et le questionnaire Blaise; il obtient ainsi un vrai fichier provisoire. Le générateur de fichiers provisoires permet donc à l'utilisateur de créer des fichiers provisoires pour les progiciels de son choix.

## 10. CONCLUSION

L'arrivée du micro-ordinateur a eu une incidence considérable sur les opérations des organismes nationaux de statistique. Le statisticien spécialisé recourt de plus en plus au micro-ordinateur et a besoin, pour son travail, d'un système intégré de traitement des données d'enquête comme celui fondé sur le système Blaise. Ce dernier se distingue d'une part par la cohérence qu'il introduit dans les diverses étapes de la collecte et du traitement des données. Il est ainsi plus facile de gérer et de contrôler le processus dans son entier. D'autre part, le système Blaise favorise l'uniformisation entre les diverses divisions. Comme toutes les divisions utilisent le même logiciel pour leurs enquêtes, l'échange de renseignements se fait plus facilement et il y a un moins grand risque d'erreur.

Le système intégré de traitement des données d'enquête est conçu plus spécialement pour des organisations très centralisées, comme le Bureau central de la statistique des Pays-Bas. Dans une organisation de cette envergure, le système intégré de traitement peut amener un accroissement substantiel de l'efficacité. Or, les organismes statistiques n'ont pas tous une structure centralisée. Dans les grands pays en particulier, le traitement des données est souvent décentralisé. Les bureaux régionaux s'occupent du traitement des données dans leurs régions respectives et font parvenir ensuite les fichiers constitués à l'administration centrale. Celle-ci combine les fichiers régionaux en un seul fichier national. Le système intégré de traitement des données d'enquête peut trouver sa place dans une telle structure: l'administration centrale élabore le questionnaire Blaise et envoie à chaque bureau régional des exemplaires du programme de saisie des données ainsi élaboré. De cette manière, on assure la cohérence des opérations de collecte et de contrôle des données dans les diverses régions. Comme les fichiers régionaux auront tous le même format (Blaise), il sera facile de les combiner en un seul fichier national en se servant des fonctions du système Blaise. De plus, comme ces fichiers seront tous "épursés", il ne sera pas nécessaire d'exécuter d'autres contrôles au niveau national. L'administration centrale n'aura plus qu'à analyser les données et les mettre en tableau et à publier les résultats de ses analyses. En outre, les bureaux régionaux ou l'administration centrale pourront publier des données régionales.

Le système Blaise est en usage depuis le milieu de 1986 et sert à un très grand nombre d'enquêtes. On le perfectionne en étroite collaboration avec les utilisateurs et chaque nouvelle version renferme des améliorations. Abacus est en usage depuis plus d'un an et est devenu très populaire auprès des utilisateurs. Quant au programme Bascula, il est encore en voie d'élaboration. On vient juste de lancer le premier prototype.



Tableau 4  
Exemple d'un tableau à deux dimensions

Nombre d'enregistrements	Population de Samplopie			
	Total	Emploi		Sexe
		Oui	Non	
Total	1,000	341	659	489
Agria	293	121	172	148
Wheaton	144	60	84	74
Greenham	94	38	56	50
Newbay	55	23	32	24
Induston	707	220	487	341
Oakdale	61	26	35	25
Crowdon	244	73	171	116
Smokeley	147	49	98	67
Mudwater	255	72	183	133

Source: Bureau statistique de Samplopie.

Le tableau ci-dessus contient de simples fréquences, mais Abacus peut calculer aussi des totaux de variables quantitatives et produire des tableaux de pourcentages et des tableaux de moyennes. De plus, les cases d'un tableau produit par Abacus peuvent contenir plus d'un élément à la fois (de fait, elles peuvent en contenir jusqu'à 10). L'utilisateur doit alors décider de la présentation du tableau (c.-à-d. classer les éléments par ligne, par colonne ou par couche). Si des données ont été recueillies au moyen d'une enquête par sondage, Abacus peut être appliqué à des données pondérées; il peut s'agir, par exemple, de poids calculés au moyen du programme Bascula. L'utilisateur n'a plus qu'à définir la variable qui contient les poids. On accorde beaucoup d'attention à la présentation du tableau car les tableaux produits par Abacus doivent être prêts à la photo. Abacus renferme donc de nombreuses options qui permettent de déterminer la présentation. Ainsi, on peut définir jusqu'à 10 lignes de texte pour l'en-tête et le cartouche d'un tableau et choisir soit des lignes horizontales et verticales (comme dans l'exemple ci-dessus), soit des lignes horizontales seulement ou ni l'une ni l'autre. La disposition du texte dans les en-têtes de colonne et la largeur des colonnes peuvent aussi être modifiées.

On peut procéder à un arrondissement afin de conserver le caractère confidentiel des données du tableau. Non seulement les totaux de case, mais aussi les totaux marginaux sont arrondis au multiple d'une constante donnée, par ex.: 5. Abacus exécute aussi bien l'arrondissement normal que l'arrondissement aléatoire. Si l'utilisateur n'est pas satisfait du tableau obtenu, il peut importer le produit d'Abacus dans le tableau Lotus 123 et pousser plus loin le traitement. Une dernière option qui mérite d'être soulignée ici est la possibilité de créer de nouvelles variables en récrivant les variables existantes (par ex.: âge groupe d'âge). Pour plus de détails sur le programme Abacus, voir Bethlehem et coll. (1989a).

## 9. ANALYSE

Le BCS n'a pas élaboré de logiciel pour l'analyse des données statistiques principalement parce que le marché compte suffisamment de bons logiciels statistiques. Le BCS utilise lui-même les logiciels SPSS (sur gros ordinateur et sur micro-ordinateur) et Stata (sur micro-ordinateur). Pour que ces logiciels puissent faire partie du système intégré de traitement des

ajustement proportionnel itératif (aussi appelé pondération multiplicative ou méthode itérative du quotient). Les poids ainsi calculés peuvent soit être ajoutés au fichier de données ou être stockés dans un fichier à part.

Bascula peut lire directement les fichiers Blaise et extraire de la spécification Blaise l'information pertinente sur les variables. Il revient à l'utilisateur de fournir l'information sur la population. Le programme est à base de menus, donc convivial. Il exécutera une stratification à posteriori complète si cela est possible. Sinon, l'utilisateur doit choisir entre une pondération linéaire ou une pondération multiplicative.

À l'heure actuelle, Bascula ne peut être utilisé que sur un micro-ordinateur. Nous prévoyons mettre au point une version "dorsale" qui tournera sur un gros ordinateur. Bascula a été conçu spécialement pour les enquêtes sociales et démographiques, où la stratification a posteriori se mêle à des méthodes d'estimation relativement simples. En ce qui concerne les enquêtes économi-ques, il faudra élaborer un logiciel différent qui mettra l'accent sur l'échantillonnage stratifié et des estimateurs plus complexes (estimateur par quotient; estimateur par régression).

## 8. ABACUS

La totalisation est une des principales étapes de la production de données statistiques et elle a été parmi les premières à être informatisées. Il existe déjà de nombreux logiciels de totalisation dans le monde mais peu sont vraiment conviviaux. Cela s'explique en partie par le fait qu'il faille définir un grand nombre de paramètres pour produire un tableau complexe selon les normes; ces paramètres peuvent être: les variables à utiliser pour les diverses dimensions (ligne, colonne, couche); le choix entre l'enchaînement des variables (c.-à-d. présenter toutes les valeurs d'une variable, puis toutes celles d'une autre variable) ou leur emboîtement (pour chaque valeur d'une variable, présenter toutes les valeurs possibles d'une autre variable) pour une dimension donnée; la quantité figurant dans les cases (fréquence, pourcentage, total, moyenne); le choix de présenter ou non les totaux et les sous-totaux; et de nombreuses caractéristiques de présentation. Pour pouvoir maîtriser tous ces paramètres, les logiciels classiques renferment des langages de commande pour définir les tableaux et ces langages ne sont pas très faciles à apprendre ni à utiliser.

Abacus est un logiciel de totalisation qui tourne sur des micro-ordinateurs dont le système d'exploitation est MS-DOS. Bien qu'on puisse le voir comme un logiciel de totalisation parmi tant d'autres, Abacus a été élaboré de façon très particulière. Premièrement, les tableaux ne sont pas définis au moyen de langages de commande. Le programme est plutôt à base de menus. L'utilisateur conçoit son tableau d'une manière simple et intuitive en mode interactif sans avoir à connaître aucun langage de commande. Deuxièmement, Abacus peut lire directement les fichiers créés par le système Blaise, de même que les fichiers ASCII. Les métadonnées, c.-à-d. l'information relative aux variables dans le fichier, peuvent être produites par le système Blaise ou encore (dans le cas de fichiers ASCII indépendants) être introduites en mode interactif par l'utilisateur. Troisièmement, le programme peut produire des tableaux prêts à la photo.

Une autre caractéristique remarquable d'Abacus est sa vitesse. Il prend environ 6 secondes pour produire un tableau que SPSS produit en 3 minutes (sur le même micro-ordinateur de type 386SX). Si Abacus est aussi rapide, c'est à cause de sa taille relativement modeste; il peut donc utiliser une grande partie de la mémoire comme zone de travail, et ainsi produire des tableaux pouvant contenir jusqu'à 90,000 cases.

Les tableaux générés par Abacus peuvent avoir jusqu'à trois dimensions (couches, lignes et colonnes). Le nombre de variables pour chaque dimension peut atteindre 10; ces variables peuvent être emboîtées ou enchaînées. Dans le tableau 4 par exemple, les variables "emploi" et "sexe" (colonnes) sont présentées selon le mode enchaînement tandis que les variables "région" et "ville" (lignes) sont présentées selon le mode emboîtement. Ce tableau ne comporte pas une troisième dimension (couches).



La première partie du questionnaire est désignée par le sigle QUEST; cette section contient la définition de toutes les questions qui peuvent être posées. Une définition se compose d'un symbole d'identification (pour usage interne dans le questionnaire), du libellé de la question telle qu'elle est posée au répondant et d'une énumération des réponses possibles. La deuxième partie du questionnaire correspond à la section CHEMINEMENT. Cette section indique dans quel ordre les questions doivent être posées et à quelles conditions. La section CONTRÔLE sert à définir les vérifications de cohérence.

La description ci-dessus ne reflète pas toute la puissance du langage Blaise. Pour avoir une idée générale de ce langage, prière de se référer à Bethlehém et coll. (1989b) et pour plus de détails, voir Bethlehém et coll. (1989c).

Le système Blaise comprend un module pour le *codage interactif*; ce module permet d'intégrer des opérations de codage à l'étape de la collecte des données ou à l'étape de la saisie et du contrôle. Le module se compose de deux "fonctions". La première met en application une approche hiérarchique du codage. Pour coder une réponse, l'opérateur commence par entrer le premier chiffre du code en choisissant la catégorie pertinente dans un menu. Une fois le premier chiffre saisi, le programme affiche un nouveau menu qui est une description plus détaillée de la catégorie choisie précédemment. Ainsi, on obtient une description de plus en plus détaillée à mesure qu'on approche du dernier chiffre. La seconde "fonction" du module représente une approche alphabétique du codage. Elle vise à trouver une description dans une liste alphabétique. Si cette description n'est pas repérée, le programme affiche la liste à partir de la description la plus près possible de la description recherchée. La liste peut être dressée de manière que l'on puisse y trouver presque toutes les descriptions possibles, y compris les permutations. Cette méthode a l'avantage d'être simple, rapide et contrôlable. Les deux fonctions de codage peuvent être appliquées simultanément.

## 7. BASCULA

Une fois "épuré", le fichier de données d'enquête produit par le système Blaise n'est pas encore prêt à servir à des inférences concernant la population d'où a été tiré l'échantillon parce que les données disponibles ne constituent pas un échantillon représentatif; il faut donc procéder à certains ajustements.

Afin de tenir compte de la non-réponse et de l'inégalité des probabilités de sélection, il faut souvent calculer des facteurs de pondération. La stratification a posteriori est une méthode de pondération bien connue. Chaque enregistré reçoit un poids calculé de telle manière que la distribution empirique pondérée de caractères tels le sexe, l'âge, l'état matrimonial et la région reflète la distribution (connue) des mêmes caractères dans la population. Deux facteurs importants peuvent compliquer la stratification a posteriori: l'existence de strates vides et le manque d'information pertinente sur la population. Des recherches ont été faites au BCS afin d'améliorer les méthodes de pondération. Ces recherches ont abouti à la mise en oeuvre d'une nouvelle méthode générale de pondération qui permet de calculer les poids à l'aide d'un modèle linéaire qui établit un rapport entre les variables étudiées dans une enquête et les variables auxiliaires. La stratification a posteriori est un cas particulier de cette méthode. En raison de la généralité de la méthode, on peut appliquer divers schémas de pondération qui exploitent le plus possible l'information existante sur la population sans soulever les problèmes mentionnés plus haut. Voir Bethlehém et Keller (1987) pour plus de détails.

Bascula est un programme général de pondération qui tourne sur des micro-ordinateurs dont le système d'exploitation est MS-DOS. Ce programme réunit plusieurs méthodes de pondération. En premier lieu, il peut exécuter une stratification a posteriori classique. Et si le nombre de strates vides est faible, il peut fonder (c'est-à-dire combiner) ces strates avec des strates avoisinantes. Lorsque le nombre de strates vides est élevé ou qu'il manque d'informations sur la population, Bascula peut exécuter la pondération linéaire décrite plus haut ou effectuer un



Tableau 3

Exemple d'un questionnaire Blaise

QUESTIONNAIRE de "l'enquête sur le travail";	
QUEST	
NumOrd	"Numéro d'ordre de l'interview?"; 1..1000 (CLÉ);
Age	"Quel âge avez-vous?"; 0..99;
Sex	"Êtes-vous de sexe masculin ou féminin?"; (Masculin, Féminin);
EtaMatr	(Marié, "Marié", "Quel est votre état matrimonial?";
Emp	NonMar "Non marié", "Occupez-vous un emploi?"; (Oui, Non);
DesEmp	"De quel genre d'emploi s'agit-il?"; [20] CARACTÈRES;
Revenu	"Quel est votre revenu annuel?";
Deplac	(Inf20 "Moins de 20,000",
	Jusq40 "Entre 20,000 et 40,000",
	Plus40 "Plus de 40,000");
	"Quel moyen de transport utilisez-vous habituellement pour aller au travail?";
	SÉRIE [3] DE
AutMoyen	(Pied "à pieds",
	Bicycle "à bicyclette",
	Aut "en voiture ou en motocyclette",
	TransCom "autobus, tramway, train ou métro",
	Autre "aucun de ceux-là");
CHEMINEMENT	
NumOrd; Age; Sex; EtaMatr; Emp;	
SI Emp = Oui ALORS	
DesEmp; Revenu; Deplac;	
SI Autre sous Deplac ALORS AutMoyen FIN	
FIN	
CONTRÔLE	
SI Age < 15 "Répondant n'a pas encore 15 ans" ALORS	
EtaMatr = NonMar "trop jeune pour être marié!"	
FIN	
FINQUESTIONNAIRE.	

programme de mise en forme de texte et place le curseur à l'endroit où se trouve approximativement l'erreur. Après la correction des erreurs, on soumet la spécification à un nouveau contrôle. Si aucune autre erreur n'est détectée, la spécification est transformée en code source Pascal, lequel est compilé en un programme exécutable.

Le langage Blaise a un double objet quelque peu contradictoire. D'une part, il doit être assez puissant pour pouvoir être appliqué à toutes les grandes enquêtes complexes et d'autre part, les spécifications de questionnaires en langage Blaise doivent être suffisamment lisibles pour que les agents spécialisés puissent s'en servir. De fait, l'information contenue dans un questionnaire Blaise doit être explicite; autrement dit, cette information représente la description fondamentale de l'enquête et est utilisable par tous les intéressés. Le tableau 3 reproduit un questionnaire simple en langage Blaise.

Enfin, on préparera une publication à l'aide du logiciel de traitement de texte PC-Write. Comme ce logiciel peut tourner sur le même système que les autres logiciels, il est facile d'importer dans le texte des tableaux et des résultats d'analyses statistiques.

## 6. LE SYSTEME BLAISE

Le système Blaise a été élaboré par le BCS et il doit son nom au célèbre théologien et mathématicien français Blaise Pascal (1623-1662). La base de ce système est le langage Blaise, qui sert à spécifier de façon formelle la structure et le contenu du questionnaire d'enquête. Le langage Blaise tire son origine, en majeure partie, du langage de programmation Pascal. Le système Blaise tourne sur des micro-ordinateurs (ou des réseaux de micro-ordinateurs) dont le système d'exploitation est MS-DOS. Il est le pivot du système intégré de traitement des données d'enquête et comme il est destiné à l'usage des employés des divisions spécialisées, nul n'a besoin d'être informaticien pour l'utiliser. On a conçu le système Blaise dans le but d'offrir aux spécialistes des divisions un outil puissant mais convivial qui leur permette de saisir les données relatives à une enquête et de voir à l'exécution de toutes les étapes subséquentes. Selon le système Blaise, la réalisation d'une enquête débute par l'élaboration d'un questionnaire en langage Blaise. Un questionnaire élaboré de cette façon contient plus de renseignements qu'un questionnaire classique. On y décrit non seulement les questions, les réponses possibles et les conditions de cheminement dans le questionnaire, mais aussi les relations entre les réponses qui doivent faire l'objet d'un contrôle.

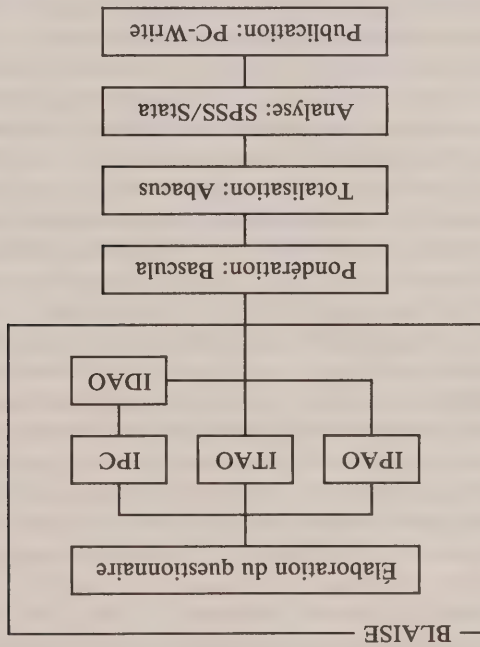
Le système Blaise peut générer des programmes pour l'IDAO ainsi que les IPAO et les ITAO. Un programme IDAO est un système intelligent et interactif destiné à la saisie et au contrôle des données recueillies à l'aide de formulaires. L'agent de la division spécialisée traite un à un les formulaires au moyen d'un micro-ordinateur. Il inscrit les réponses à l'endroit approprié et une fois le questionnaire terminé, il actionne la fonction contrôle pour vérifier s'il n'y aurait pas d'erreur de cheminement ou d'erreur de logique. Les erreurs détectées sont affichées à l'écran avec une explication. L'agent peut corriger ces erreurs soit en se reportant au formulaire ou bien en appelant la personne qui a fourni les renseignements en question. Une fois que toutes les erreurs ont été corrigées, l'enregistrement "épuré" est versé dans un fichier.

Le programme IDAO peut aussi servir à une forme de traitement de données dont on n'a pas fait mention jusqu'à maintenant. En effet, les organismes statistiques ne recueillent pas toujours eux-mêmes les données dont ils ont besoin; ils sont appelés parfois à construire des statistiques à l'aide de fichiers provenant de l'extérieur. Ces données doivent tout de même être vérifiées. Le système Blaise comporte une fonction qui permet d'importer les fichiers de ce genre. Grâce au programme IDAO, on peut soumettre tous les enregistrements à un contrôle intégral par lots. Les enregistrements sont alors classés en deux groupes: "sans erreur" et "avec erreur". Les seconds peuvent être corrigés en mode interactif grâce, une fois de plus, au programme IDAO.

Les programmes IPAO et ITAO peuvent servir aux interviews assistées par ordinateur. Dans ce cas, le questionnaire classique est remplacé par un programme informatique qui contient les questions à poser. Ce programme contrôle entièrement l'interview. Il détermine au fur et à mesure l'ordre dans lequel les questions doivent être posées et vérifie les réponses dès qu'elles ont été saisies. En ce qui concerne l'IPAO, le programme est chargé dans un ordinateur portable que l'intervieweur apporte au domicile du répondant. Pour ce qui a trait à l'ITAO, le programme est chargé dans un ordinateur de bureau. L'intervieweur appelle le répondant depuis une unité centrale et réalise l'interview par téléphone.

La génération d'un programme IDAO, IPAO ou ITAO de Blaise se fait en un certain nombre d'étapes. Premièrement, on introduit la spécification en langage Blaise du questionnaire au moyen d'un programme de mise en forme de texte; on vérifie ensuite s'il n'y a pas d'erreur de syntaxe. Les erreurs détectées doivent être corrigées et à cette fin, le système rappelle le

Tableau 2  
Système intégré de traitement des données d'enquête



logiciels de traitement de données (par ex.: logiciels pour la totalisation et l'analyse). Ainsi, la répétition des spécifications relatives aux données n'a plus raison d'être et la cohérence des données est observée à toutes les étapes du traitement.

Le pivot du système intégré de traitement des données d'enquête mis au point par le BCS est le système Blaise. À l'étape de la conception de l'enquête, le questionnaire est spécifié en langage Blaise et c'est cette spécification qui sert, durant tout le processus, à extraire les données requises pour les diverses étapes du traitement. Le tableau 2 décrit schématiquement le système intégré de traitement des données d'enquête.

Le système Blaise peut générer trois types de programmes: IDAO, IPAO et ITAO (le sigle IDAO signifie "introduction de données assistée par ordinateur"). Ce système intègre les opérations de saisie et de contrôle des données en permettant le traitement interactif des questionnaires. Il peut aussi produire le logiciel nécessaire pour les IPAO ou les ITAO. Le système Blaise est décrit plus en détail dans la section 6.

Quel que soit le mode de collecte des données, nous obtiendrons toujours un fichier "épuré", c.-à-d. un fichier où on ne peut plus trouver d'erreurs. Viendra ensuite le calcul de facteurs de pondération. Le programme Bascula est conçu spécialement à cette fin. Il peut lire directe-ment les fichiers Blaise et extraire de la spécification Blaise l'information relative aux variables, c'est-à-dire les métadonnées. L'exécution de Bascula aura pour effet d'ajouter une variable dans le fichier contenant les facteurs de pondération pour les divers cas. Le programme Bascula est décrit plus en détail dans la section 7.

Le fichier est maintenant prêt pour la totalisation et à cette fin, le système intégré met à notre disposition le programme Abacus. Ce programme peut, lui aussi, lire et comprendre les fichiers créés à l'étape précédente. Voir la section 8 pour plus de détails. La totalisation peut être suivie d'une analyse plus approfondie des données. À cette fin, le système Blaise peut générer des inter-faces pour les logiciels statistiques SPSS et Stata. La section 9 contient plus de détails à ce sujet.



De plus, les utilisateurs dans les divisions spécialisées négligent souvent d'effectuer des opérations comme la sauvegarde de fichiers et l'archivage. Par ailleurs, les divisions ne peuvent se partager l'utilisation de fichiers autrement que par l'échange de disquettes. Enfin, pour les grandes organisations qui possèdent beaucoup de micro-ordinateurs autonomes, la distribution des nouvelles versions de logiciels, y compris la documentation pertinente, est une tâche souvent contraignante.

Pour éviter toutes ces difficultés, le BCS a installé quelque 60 réseaux locaux (RL). Chaque division a son propre RL. De dix à soixante micro-ordinateurs sont reliés à un serveur de fichiers haut de gamme articulé sur un microprocesseur 386 et dont la capacité de mémoire peut atteindre 600 mégaoctets. Ainsi, près de 2,300 micro-ordinateurs sont reliés de cette manière à un serveur de fichiers et la moitié d'entre eux sont articulés autour du microprocesseur 386SX d'Intel. La sécurité est assurée de trois façons : – protection par mot de passe dans la procédure d'entrée en communication, – chiffrage et – postes de travail sans unité de disquette (seulement 60 des 2,300 micro-ordinateurs sont équipés d'une unité de disquette ou de disque dur). Les opérations d'archivage et de sauvegarde des RL s'effectuent de façon centralisée sous la responsabilité de la division de l'informatique. Chaque nuit, on procède à une sauvegarde complète de plus de 15 giga-octets. Il est clair qu'un tel contexte d'utilisation facilite des opérations comme le contrôle de la version et la mise à jour du logiciel. Avec un RL, la distribution et l'installation des nouvelles versions de logiciels deviennent des tâches simples par une seule commande, on peut télécharger de nouvelles versions dans tous les serveurs de fichiers à la fois. Tous les permis d'utilisation de logiciel prévoient un usage en commun, qui est contrôlé par des logiciels "maison".

Le micro-ordinateur joue un rôle de plus en plus grand dans la production de données statistiques, mais le gros ordinateur ou le mini-ordinateur sont encore indispensables pour certaines applications, comme l'exploitation de grosses bases de données. Dans ce contexte, le BCS a fait de Oracle son système de gestion de bases de données standard. L'élaboration d'un programme de gestion de bases de données se fait de préférence sur un micro-ordinateur alors que l'exécution de ce programme se fait sur un mini-ordinateur. Dernièrement, le BCS a élaboré une architecture client-serveur fondée sur un système de gestion de base de données réparties. Les micro-ordinateurs du réseau constituent le système frontal et les mini-ordinateurs, le système dorsal.

Ainsi, plus les outils informatiques sont à la portée des agents spécialisés des divisions (décentralisation), plus il est nécessaire de centraliser les opérations d'uniformisation et de coordination de l'univers informatique des utilisateurs. Pour plus de détails sur l'infrastructure informatique, voir Keller, Metz et Bethlehem (1990).

## 5. INTÉGRATION DU TRAITEMENT DES DONNÉES D'ENQUÊTE

Dans la section précédente, nous avons discuté de la nécessité d'une intégration du traitement des données d'enquête. Il a été question plus particulièrement de la concentration des tâches dans les divisions spécialisées et de l'uniformisation du matériel et du logiciel. Or, l'uniformisation du logiciel ne suffit pas. On peut accroître davantage l'efficacité de la production de données statistiques en intégrant les logiciels nécessaires dans un système. Cette section vise à décrire la manière dont le BCS a mis en oeuvre un système intégré de traitement des données d'enquête.

Un système intégré de traitement des données d'enquête doit reposer sur un langage puissant pour la spécification des questionnaires. Cette spécification est la "base de connaissances", qui contient tous les faits relatifs au questionnaire et aux données. Le système intégré doit pouvoir exploiter ces connaissances, c'est-à-dire qu'il doit pouvoir générer automatiquement toutes les applications requises pour le traitement des données, soit d'une part, les logiciels pour la collecte, la saisie et le contrôle des données et d'autre part, les interfaces avec d'autres

il va de soi que l'on opte pour le micro-ordinateur. On peut ainsi s'offrir la convivialité à un prix relativement raisonnable et avoir accès à toute une série de logiciels utiles. Sachant que les organismes statistiques traitent régulièrement d'énormes quantités de données, on peut se demander si les micro-ordinateurs peuvent à eux seuls suffire à la tâche et remplacer les gros ordinateurs. Pour être en mesure de répondre à cette question, nous devons distinguer deux types d'opérations liées au traitement des données. Premièrement, les opérations axées sur l'enregistrement, c'est-à-dire celles qui portent sur un seul enregistrement à la fois (par ex. : saisie et contrôle des données). Les opérations de ce genre se prêtent habituellement très bien au traitement interactif. Deuxièmement, les opérations axées sur le fichier, c'est-à-dire celles qui ne peuvent être exécutées convenablement que si l'on dispose de tout le fichier (par ex. : calcul des poids et totalisation). En raison de leur envergure, les opérations axées sur le fichier sont souvent exécutées suivant le principe du traitement par lots.

Il y a quelques années, on était d'avis que les opérations axées sur l'enregistrement pouvaient être faites sur micro-ordinateur mais que celles axées sur le fichier devaient être exécutées sur un gros ordinateur. Comme on fabrique des micro-ordinateurs de plus en plus puissants, l'attention est tournée de plus en plus vers le micro-ordinateur. À l'heure actuelle, la position du BCS est d'utiliser le micro-ordinateur pour toutes les opérations axées sur l'enregistrement et pour un grand nombre des opérations axées sur le fichier (pour des fichiers de moins de 50 mégaoctets par exemple). En revanche, nous conservons le gros ordinateur pour le stockage des données et les gros travaux en traitement par lots.

Les personnes qui doivent utiliser un ordinateur devraient être le moins possible mises en contact avec un gros ordinateur. C'est pourquoi le BCS favorise de plus en plus l'utilisation de systèmes frontal et dorsal. Le système frontal est constitué de micro-ordinateurs et c'est ce qu'utilisent les statisticiens pour traiter les problèmes courants. Le système dorsal consiste en un gros ordinateur ou en un mini-ordinateur et sert aux gros travaux; les opérations pertinentes peuvent même se faire à l'insu de l'utilisateur. En particulier, le serveur de base de données est une option très intéressante en ce qui concerne la gestion de bases de données. Selon cette option, les opérations portant sur la base de données s'effectuent sur un mini-ordinateur spécialisé tandis qu'elles sont spécifiées, déclenchées et contrôlées au moyen d'un micro-ordinateur placé sur le bureau de l'utilisateur.

#### 4. UNIFORMISATION

Le BCS utilise de plus en plus le micro-ordinateur (avec système MS-DOS) à beaucoup d'étapes de la production de données statistiques. D'une part cette initiative favorise un traitement plus efficace de l'information, de l'autre elle crée de nouveaux problèmes qu'il faut chercher à résoudre. Si on laisse à chaque division la liberté d'acheter le matériel et le logiciel au plus de deux logiciels pour une tâche particulière.

Un autre avantage de l'uniformisation est qu'elle limite le nombre de cours de formation que doivent suivre les utilisateurs. Afin d'assurer la formation de nombreux nouveaux utilisateurs, le BCS organise en moyenne chaque mois 50 cours d'une journée (ce qui suffit à occuper trois salles de cours, dotées de tout le matériel nécessaire, à chaque jour ouvrable).

On doit aussi porter attention à la manière dont sont utilisés les micro-ordinateurs dans l'organisation. La distribution d'un grand nombre de micro-ordinateurs autonomes peut sembler une solution commode mais cela soulève des problèmes qu'il faut chercher à résoudre. Par exemple, il est très facile de copier des fichiers de données (confidentielles) sur le disque dur d'un micro-ordinateur autonome; il y a donc là un problème de sécurité informatique.



## - Multiplicité des étapes

L'information passe entre plusieurs mains: le répondant remplit le questionnaire, l'agent spécialisé en vérifie le contenu et corrige les erreurs, l'agent de saisie introduit les données dans l'ordinateur et le programmeur élabore des logiciels de contrôle. L'échange de documents entre des personnes ou des divisions peut être une source d'erreur, de confusion et de retard.

## - Multiplicité des systèmes informatiques

Les diverses opérations liées au traitement des données peuvent ne pas être exécutées sur le même système. Le transfert de fichiers provoque des retards; en outre, une spécification et une documentation fautives peuvent être à l'origine d'erreurs.

## - Répétition des spécifications relatives aux données

La structure des données doit être spécifiée à presque toutes les étapes. Les questions qui reviennent le plus souvent dans les divisions sont les suivantes: Quelle est la signification des variables? Quelles valeurs peuvent-elles prendre? Des conditions s'appliquent-elles au processus de cheminement? Quelles relations entre les variables faut-il vérifier? Bien que les spécifications soient essentiellement les mêmes d'une fois à l'autre, le mode de spécification, lui, peut être très différent. Chaque système a son propre "langage". La première spécification concerne le questionnaire proprement dit. Une deuxième spécification peut être nécessaire pour la saisie des données, et une autre pour le programme de contrôle, la totalisation et l'analyse, etc. De toute évidence, ce n'est pas la façon la plus efficace de traiter l'information.

Devant ces difficultés, le BCS propose comme solution l'*intégration*, une intégration sur trois plans: tâches, matériel et logiciel. Examinons tout d'abord l'intégration des tâches.

Dans sa forme classique, le traitement des données consiste en ce que nous appelons des *macro-cycles*. L'ensemble des données d'une enquête passe par des cycles: d'une division à une autre et d'un système informatique à un autre. En premier lieu, la division spécialisée s'occupe d'"épurer" manuellement les questionnaires; ensuite, la section de la saisie des données voit à l'introduction des données de ces questionnaires; en troisième lieu, les fichiers sont transférés dans un gros ordinateur. Un programme soumet les données à une vérification de cohérence; les erreurs détectées sont inscrites sur une liste que l'on envoie à la division spécialisée pour qu'elle fasse les corrections nécessaires. Ce processus de saisie et de contrôle doit être répété un certain nombre de fois avant que l'on puisse dire que les données sont "épures".

Ce que vise au fond l'intégration des tâches est le remplacement des macro-cycles par des *micro-cycles*. Autrement dit, on devrait songer à faire circuler un seul enregistrement à la fois plutôt que tout le fichier de données. La notion de micro-cycle signifie que toutes les opérations successives se font à l'intérieur du même système informatique et qu'elles sont sous le contrôle d'une seule division. Substituer le micro-cycle au macro-cycle revient à concentrer toutes les opérations de traitement des données dans une seule division, soit la division spécialisée. Comme les statisticiens des divisions spécialisées sont les personnes qui connaissent le mieux le domaine sur lequel porte une enquête, ils sont le plus en mesure de traiter les données, de résoudre les problèmes et de produire des statistiques de première qualité. Pour s'acquitter de ces tâches, il faut, bien sûr, qu'ils aient les outils appropriés, c.-à-d. des systèmes puissants et conviviaux.

L'idée de confier les opérations de traitement des données d'enquête à des spécialistes de l'informatique est révolue. Les statisticiens des divisions spécialisées se rendent compte de plus en plus des possibilités et de l'utilité de l'ordinateur pour leur travail. C'est pourquoi les divisions spécialisées devraient songer à prendre en charge les opérations liées au traitement des données d'enquête, dans la mesure où il s'agit d'opérations peu complexes. Evidemment, il revient à la division de l'informatique de mettre en place l'infrastructure nécessaire. Par ailleurs, cette division demeure responsable de l'élaboration et de la maintenance des systèmes d'information complexes.

Le second volet de la solution du BCS est l'intégration du matériel. Celle-ci vise à faire exécuter le maximum d'opérations sur un seul type d'ordinateur. Compte tenu de ce qu'un grand nombre de statisticiens peu initiés à l'informatique devront se servir d'un ordinateur,



composition des lignes et des colonnes (construites souvent à partir de plusieurs variables), les quantités figurant dans les cases (effectifs, moyennes, pourcentages), le mode de calcul des pourcentages, le traitement des variables à réponse multiple, la position des totaux et des sous-totaux et beaucoup d'autres facteurs peuvent compliquer grandement la tâche des statisticiens.

De nombreux organismes statistiques soumettent aussi leurs données à des analyses qui visent à en révéler la structure fondamentale; ils peuvent ainsi faire une meilleure lecture de ces données. Les résultats de ces analyses peuvent contribuer à perfectionner les enquêtes ultérieures et par conséquent à améliorer la qualité des données ou à réduire les coûts des opérations.

Enfin, les résultats des analyses sont *publiés* sous forme de rapport. Ce rapport renferme normalement des tableaux et des graphiques. Il est essentiel de présenter l'information statistique de manière à ce qu'elle soit comprise de la bonne façon. Les graphiques, les tableaux et les textes doivent être simples et limpides. La construction des graphiques mérite une attention particulière car un graphique équivoque donnera lieu très facilement à de fausses interprétations.

### 3. NÉCESSITÉ DE L'INTÉGRATION

L'ordinateur a toujours occupé une place importante dans le traitement des données statistiques. À l'origine, il servait uniquement à des opérations comme le tri, le dénombrement et la totalisation. Dans les années soixante et soixante-dix, l'apparition des gros ordinateurs et des progiciels statistiques a permis l'exécution d'analyses plus poussées. L'ordinateur servait aussi de plus en plus au contrôle des données, à la pondération et à l'imputation. Ce n'est que plus tard que l'ordinateur commença à servir à la collecte des données. La première occasion sera les interviews téléphoniques (ITAO). La dernière décennie a vu naître l'ordinateur portable; désormais, l'intervieweur peut se rendre au domicile du répondant avec un ordinateur sous le bras. Cette forme d'interview est désignée par le sigle IPAO (interview sur place assistée par ordinateur).

Il ne fait plus de doute que l'ordinateur est utilisé à des fins de plus en plus nombreuses. Il existe du matériel et des logiciels pour presque toutes les étapes du processus de production. En outre, de plus en plus de gens utilisent les outils informatiques. Au début, les ordinateurs étaient réservés à l'usage des spécialistes de l'informatique mais aujourd'hui, les statisticiens et les experts des divisions spécialisées sont rompus aux techniques de l'informatique et travaillent avec toujours plus d'intérêt le matériel et les logiciels voulus pour accomplir leur tâche. À l'heure actuelle, le traitement électronique des données peut être confié aux divisions spécialisées pourvu qu'il n'implique pas d'opérations complexes; de fait, les divisions se chargent déjà de cette opération, laissant aux spécialistes de la division de l'informatique l'élaboration et la maintenance des systèmes d'information complexes. Ces changements ont eu pour conséquence de modifier le travail des statisticiens et des experts des divisions spécialisées. Auparavant, ces gens étaient des spécialistes de leur domaine propre, sans plus; mais aujourd'hui, ils ont élargi leur champ de connaissances et diversifié leur expérience en touchant aussi bien à la méthodologie statistique et à l'informatique qu'à des sujets spécialisés. Ainsi, le groupe de spécialistes a fait place à un nouveau groupe qui est plus au fait des divers aspects du traitement des données d'enquête.

Il est bien d'automatiser la production de données statistiques mais il faut être conscient des risques que cela implique. Même si l'utilisation de l'ordinateur est le gage d'une plus grande efficacité et d'une meilleure qualité des données, un usage incontrôlé des nouvelles techniques peut facilement mener à la confusion et, partant, à une productivité moindre. Voici les facteurs qui influent sur l'efficacité de la production de données statistiques:

Si la collecte des données se fait au moyen de questionnaires, ceux-ci, une fois remplis, il est indispensable d'extraire de ces questionnaires toutes les erreurs que l'on peut y trouver. Nous en sommes à l'étape du *contrôle des données*. On y distingue trois types d'erreur : - l'*erreur d'inscription* (lorsque la réponse donnée est en dehors de l'intervalle des réponses possibles, par ex. : âge = 348 ans) ; - l'*erreur de logique* (lorsqu'il y a incohérence entre des réponses à certaines questions, par ex. : il est raisonnable de croire qu'une personne peut être âgée de 8 ans, et il n'y a rien d'anormal à ce qu'une personne se dise mariée, mais lorsque ces deux caractéristiques se rapportent au même individu, il y a sûrement quelque chose qui cloche, du moins en ce qui concerne les Pays-Bas) ; - l'*erreur de cheminement* (lorsque l'intervieweur ou le répondant ne respecte pas les instructions relatives à la manière de remplir le questionnaire, c'est-à-dire lorsqu'il ne tient pas compte de l'enchaînement logique des questions : ainsi, des réponses sont données à des questions qui ne concernent pas l'individu, et vice-versa).

Les erreurs qui ont été décelées doivent être corrigées, mais cette tâche peut s'avérer très ardue si elle doit exécutée postérieurement, dans les bureaux de l'organisme statistique. Dans beaucoup de cas, particulièrement en ce qui concerne les enquêtes-ménages, il est impossible de joindre de nouveau le répondant ; il faut donc trouver d'autres moyens de résoudre la difficulté. Parfois, on réussira à déterminer une valeur approximative acceptable à l'aide d'une méthode d'imputation mais dans d'autres circonstances, on remplacera la valeur inexacte par un code spécial indiquant que la valeur est "inconnue".

À l'étape du contrôle des données, on procède aussi parfois à ce qu'on appelle le *codage des réponses libres*. Un exemple typique de question à réponse libre est la question relative à la profession du répondant. Il est évidemment plus facile de traiter les questions qui renferment une liste de réponses pré-codées. Toutefois, dans le cas de la profession, cette liste serait interminable et le répondant aurait beaucoup de difficulté à faire le bon choix. Pour éviter ce problème, il suffit de laisser le répondant inscrire en toutes lettres sa profession sur le questionnaire. Ces réponses doivent être ensuite classées si l'on veut être en mesure d'analyser ce genre d'information. La classification des réponses est une opération longue et coûteuse qui doit être confiée à des spécialistes chevronnés.

Le contrôle des données produit un fichier "épuré", c.-à-d. un fichier ne contenant pas d'erreur. Cependant, il est encore trop tôt pour procéder à la totalisation et à l'analyse des données. En premier lieu, l'échantillonnage peut avoir été fait avec des probabilités inégales ; par exemple, des établissements ont été échantillonnés avec une probabilité proportionnelle à la taille. Un choix judicieux des probabilités d'échantillonnage permet d'obtenir des estimations plus précises des paramètres de la population mais à la condition que l'on utilise une méthode d'estimation qui tient compte des différences de probabilité. En second lieu, la représentativité de l'échantillon peut être affectée par la non-réponse, c'est-à-dire le phénomène selon lequel il manque des données pour certains éléments de l'échantillon. Si le comportement des non-répondants varie en fonction des caractéristiques démographiques étudiées, les résultats de l'enquête seront biaisés.

Afin de compenser les effets de la non-réponse et de l'inégalité des probabilités d'échantillonnage, on a souvent recours à la *pondération*. À cette étape, on attribue un poids à chaque enregistré. Les poids sont calculés de telle manière que la distribution empirique pondérée de ces caractères dans la population.

Dans les cas de non-réponse partielle, c'est-à-dire les cas où un enregistré ne contient pas toutes les réponses voulues, on peut aussi procéder à une *imputation*. En se servant d'un modèle quelconque, on calcule des estimations pour les valeurs manquantes et on les inclut dans l'enregistrement.

À l'issue de cette étape, nous avons un fichier épuré prêt pour l'*analyse*. Celle-ci débute presque toujours par la totalisation des caractéristiques de base. La construction de tableaux n'est pas une opération aussi simple qu'on pourrait le croire. En effet, des facteurs tels la



2. LA PRODUCTION DE DONNÉES STATISTIQUES

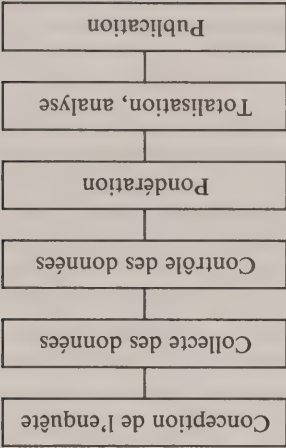
Les organismes nationaux de statistique recueillent des données sur les particuliers, les ménages et les établissements et transforment cette information en des statistiques utiles. La production de données statistiques est un processus souvent complexe, long et coûteux. Cette section vise à décrire les étapes que doit franchir l'organisme de statistique pour mener à bien une enquête, les problèmes qui peuvent se poser en cours de route ainsi que les décisions qui doivent être prises. Le tableau 1 donne un aperçu des diverses étapes d'une enquête.

La première étape est évidemment la conception de l'enquête; le statisticien doit alors déterminer quelle population sera étudiée, quelles données seront recueillies et quelles caractéristiques seront estimées. Comme les organismes statistiques recueillent la plupart de leurs données au moyen d'enquêtes par sondage, il faut élaborer un questionnaire qui contiendra toutes les questions qui seront posées aux répondants. Ce questionnaire est la première description pratique des données à recueillir. De plus, comme il s'agit de sondages, le statisticien doit aussi élaborer un plan d'échantillonnage et veiller à ce que l'échantillon soit prélevé selon les normes.

La deuxième étape du processus est la *collecte des données*. Dans de nombreuses enquêtes, les questionnaires sont remplis généralement à l'occasion d'une interview sur place: un intervieweur se rend à domicile, pose les questions et inscrit les réponses sur un formulaire. La qualité des données recueillies est généralement bonne. Cependant, comme cette méthode nécessite un grand nombre d'intervieweurs, qui peuvent tous être appelés à effectuer de nombreux déplacements, l'interview sur place est une formule qui peut s'avérer coûteuse et chronophage. C'est pourquoi on a recours parfois à l'interview téléphonique. Les intervieweurs appellent les répondants depuis les bureaux de l'organisme; ils n'ont donc plus besoin de se déplacer. En revanche, il n'est pas toujours possible de réaliser des interviews téléphoniques: seules les personnes qui ont le téléphone peuvent être rejointes de cette manière, et le questionnaire ne doit pas être trop long ni trop compliqué. L'enquête postale est une solution encore moins coûteuse car elle n'exige pas les services d'intervieweurs. On envoie un questionnaire aux répondants éventuels et ceux-ci sont priés de le retourner dûment rempli. Bien que l'on dispose de lettres de rappel dans ces conditions, cela n'équivaut pas au pouvoir de persuasion d'un intervieweur; par conséquent, le taux de réponse et la qualité des données recueillies en souffrent.

Tableau 1

La production de données statistiques





# Le système Blaise ou comment élaborer un système de traitement intégré des données d'enquête

JELKE G. BETHLEHEM et WOUTER J. KELLER<sup>1</sup>

## RÉSUMÉ

Grâce aux progrès récents de l'informatique, les organismes nationaux de statistique sont en mesure de produire efficacement des données de première qualité. Le Bureau central de la statistique des Pays-Bas (BCS) recourt de plus en plus au micro-ordinateur pour les diverses étapes de la production de données statistiques. Dans cet article, nous examinons le rôle que jouent les logiciels et les machines dans la collecte et le contrôle des données ainsi que dans la totalisation et l'analyse. Afin de sensibiliser le lecteur aux conséquences néfastes d'une gestion anarchique d'un système décentralisé de traitement de données, nous faisons ressortir toute l'importance de l'intégration. Celle-ci rend la production de données statistiques plus facile à gérer et plus efficace. Cet article est aussi une occasion de présenter le système Blaise, élaboré par le BCS, comme un outil informatique favorisant l'intégration du traitement des données d'enquête. À partir d'une description du questionnaire d'enquête, ce système peut générer automatiquement divers programmes destinés à la collecte des données (IPAO ou ITAO) ou à la saisie et au contrôle des données (IDAO). Il peut aussi être en liaison avec d'autres logiciels. À cet égard, nous décrivons le lien qui existe entre Blaise et les logiciels "maison" (Bascula pour la pondération) et Abacus (pour la totalisation). Ainsi, le système Blaise contrôle et coordonne, et par conséquent intègre, la majeure partie des opérations d'enquête.

**MOTS CLÉS:** Intégration; traitement des données d'enquête; IPAO; ITAO; micro-ordinateurs; décentralisation; uniformisation.

## 1. INTRODUCTION

Le Bureau central de la statistique des Pays-Bas (BCS) utilise de plus en plus des micro-ordinateurs pour le traitement des données d'enquête. L'arrivée du micro-ordinateur a eu des effets considérables sur les méthodes de travail de l'organisme statistique néerlandais. Les statisticiens des divisions spécialisées prennent conscience graduellement des possibilités de ce nouvel appareil et c'est pourquoi ils l'utilisent de plus en plus fréquemment dans le quotidien.

Dans cet article, nous allons examiner le rôle des nouvelles techniques informatiques dans la collecte et le contrôle des données ainsi que dans la totalisation et l'analyse. Nous allons insister sur l'importance de l'uniformisation et de l'intégration. Ces mesures présentent trois avantages concrets: elles nous épargnent les effets néfastes d'une gestion anarchique d'un système décentralisé de traitement de données, elles facilitent la gestion de la production de données statistiques et favorisent une plus grande efficacité.

Nous présentons de plus le système Blaise, une création du BCS, comme le pivot d'un système intégré de traitement des données d'enquête. D'une part, Blaise se caractérise plus particulièrement par la cohérence qu'il introduit dans les diverses étapes de la collecte et du traitement des données. D'autre part, il favorise l'uniformisation des méthodes entre les divisions. Si toutes les divisions utilisent le même logiciel pour traiter les données de leurs enquêtes, tous parleront le même "langage" et l'échange de renseignements entre divisions sera plus simple et moins sujet à l'erreur.

<sup>1</sup> Jelke G. Bethlehem et Wouter J. Keller, Bureau central de la statistique, division de l'informatique, C.P. 959, 2270 AZ Voorburg, Pays-Bas.

le nombre de numéros échantillonnés par grappe est si faible. De plus, la variance des estimations obtenues avec la méthode modifiée sera de 20 à 30% plus élevée que si on utilisait la méthode originale, car la troncation n'est pas très efficace avec les échantillons de petite taille. Par conséquent, dans la plupart des enquêtes où il y a moins de dix numéros échantillonnés par grappe, il faut que la méthode originale pose de très graves problèmes d'application avant qu'on décide de l'abandonner et d'employer la méthode modifiée.

## REMERCIEMENTS

Les auteurs tiennent à remercier les arbitres de leurs observations, qui ont été très utiles et qui ont permis de clarifier le contenu du présent article.

## BIBLIOGRAPHIE

- CUMMINGS, K. M. (1979). Random digit dialing: A sampling technique for telephone surveys. *Public Opinion Quarterly*, 233-244.
- DREW, J. D., DICK, P., et SWITZER, K. (1989). Development and testing of telephone survey methods for household surveys at Statistics Canada. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 120-127.
- GROVES, R. M., BIEMER, P. P., LYBERG, L. E., MASSEY, J. T., NICHOLS, W. L., et WAKSBERG, J. (eds.) (1988). *Telephone Survey Methodology*. New York: John Wiley and Sons.
- HANSEN, M. H., HURWITZ, W. N., et MADOW, W. G. (1953). *Sample Survey Methods and Theory*, 2. New York: John Wiley and Sons, 138-139.
- KISH, L. (1965). *Survey Sampling*. New York: John Wiley and Sons, 429-430.
- LEPKOWSKI, J. M., et GROVES, R. M. (1986). A two phase probability proportional to size design for telephone sampling. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 73-98.
- PIEKARSKI, L. (1990). Working block density declines. *The Frame*, une publication de Survey Sampling Inc.
- POTTHOFF, R. F. (1987). Generalizations of the Mitofsky-Waksberg technique for random digit dialing. *Journal of the American Statistical Association*, 82, 409-418.
- SUDMAN, S. (1973). The uses of telephone directories for survey sampling. *Journal of Marketing Research*, 10, 204-207.
- TUCKER, C. (1989). Characteristics of commercial residential telephone lists and dual frame designs. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 128-137.
- WAKSBERG, J. (1984). Efficiency of alternative methods of establishing cluster sizes in RDD sampling. Note de service non publiée de Westat Inc.
- WAKSBERG, J. (1978). Sampling methods for random digit dialing. *Journal of the American Statistical Association*, 73, 40-46.
- WAKSBERG, J. (1973). The effect of stratification with differential sampling rates on attributes of subjects of the population. *Proceedings of the Social Statistics Section, American Statistical Association*.

Il est difficile de formuler des recommandations précises quant au moment où il faut employer la version originale ou la version modifiée de la méthode de Mitofsky et Waksberg, car tout dépend de circonstances qui varient d'une enquête à l'autre. Nous proposons ci-après des lignes directrices concernant le choix de la méthode appropriée.

En règle générale, il est préférable d'utiliser la version originale de la méthode dans le cas des enquêtes nécessitant un contrôle très rigoureux de la taille de l'échantillon ou le prélèvement d'un échantillon pratiquement autopondéré. Bien que la méthode modifiée permette d'estimer avec relativement de précision la taille de l'échantillon, on peut s'attendre à une certaine variation de cette taille, surtout en raison de l'incertitude relative aux taux de non-réponse. Le prélèvement d'un échantillon autopondéré, qu'on ne peut réussir à l'aide de la version modifiée de la méthode, présente par ailleurs certains avantages découlant de la simplification de l'analyse statistique standard.

Comme les coûts d'utilisation des deux méthodes sont différents, il serait utile de disposer de modèles d'estimation de l'écart des coûts pour nous aider à les évaluer. Malheureusement, il n'est pas facile de quantifier les différences entre les coûts inhérents aux versions originale et modifiée de la méthode de Mitofsky et Waksberg. De fait, l'absence de modèles de coûts raisonnables constitue un des principaux problèmes qui restreint notre capacité d'établir des plans de sondage optimaux.

En l'absence de modèles raisonnables d'estimation de l'écart des coûts, nous proposons certaines conditions susceptibles de favoriser l'utilisation d'une méthode plutôt que de l'autre. Ainsi, il est préférable d'utiliser la version modifiée de la méthode lorsque la durée de l'interview est relativement brève. À mesure que l'interview se prolonge, les économies réalisées par suite de l'utilisation de la version modifiée de la méthode risquent de devenir moins élevées, tandis que les variances des estimations s'accroissent.

La durée de l'interview est particulièrement importante dans le cas des enquêtes où l'on soumet les ménages à une présélection afin de trouver ceux qui possèdent certaines caractéristiques. Ainsi, dans certaines enquêtes téléphoniques à composition aléatoire, on soumet les ménages à une sélection préliminaire et on n'effectue l'interview que si un des membres du ménage fait partie d'un groupe cible donné. En pareil cas, l'interview de présélection est souvent très brève. Il peut alors être très avantageux d'utiliser la version modifiée de la méthode de Mitofsky et Waksberg. Les enquêtes au cours desquelles on soumet les ménages à une présélection tendent aussi à avoir un nombre plus élevé d'unités échantillonnées par grappe, ce qui rend la version modifiée de la méthode encore plus efficace. Lorsque dix numéros ou plus sont échantillonnés par grappe (ce qui équivaut à environ six ménages par grappe), les biais dans les estimations obtenues en utilisant la version modifiée de la méthode de Mitofsky et Waksberg sont pratiquement sans importance, et l'utilisation de poids tronqués permet de réduire l'augmentation de la variance à environ 10%. Il est souvent acceptable de prélever des échantillons de dix numéros ou plus par grappe aux fins de la sélection préliminaire, bien qu'il soit généralement inefficace de prélever des échantillons de cette taille dans chaque grappe pour les fins de l'interview, même lorsque le coefficient de corrélation interne intra-classe est peu élevé.

Compte tenu de ces facteurs, on peut de façon générale recommander l'utilisation de la version modifiée de la méthode de Mitofsky et Waksberg lorsque les ménages à l'intérieur des grappes doivent être soumis à une sélection préliminaire. De façon plus explicite, on devrait considérer l'emploi de la méthode modifiée avec des poids tronqués dans les cas suivants: (1) lorsque dix numéros ou plus sont échantillonnés par grappe, et (2), lorsque le coût total de l'utilisation de la méthode modifiée est inférieur d'au moins 10% à celui de l'emploi de la version originale, ou lorsque la période de collecte des données est relativement courte. Si l'on ne satisfait pas à ces deux conditions, il faut fonder le choix de la méthode sur des évaluations relatives à d'autres exigences de l'enquête.

Lorsque le nombre de numéros échantillonnés par grappe est inférieur à dix, le biais et la variance provoqués par l'utilisation de la version modifiée de la méthode de Mitofsky et Waksberg constituent alors des problèmes plus graves. Toutes les caractéristiques corrélées avec la proportion de numéros résidentiels dans une grappe pourraient être touchées lorsque



des échantillons de dix numéros par grappe, tous les estimateurs, à l'exception de celui utilisant un poids non corrigé, sont raisonnablement efficaces. Le biais obtenu en utilisant le poids  $(n_i + .5)^{-1}$  est particulièrement encourageant.

Pour ce qui est de l'échantillon le plus petit étudié, soit celui de cinq numéros par grappe, les risques de biais sont un peu plus élevés. L'utilisation du poids naturel,  $n_i^{-1}$ , permet d'obtenir un biais dans la taille de l'échantillon un peu moins élevé que celle du poids  $((n_i + .5)^{-1})$ , mais la différence entre les deux biais n'est pas très grande. Le biais relatif dans la taille de l'échantillon qu'on obtient lorsqu'on utilise ces deux poids est toujours inférieur à 1%. Le biais est positif dans le cas de densités de numéros résidentiels qui se situent entre 45 et 80% et négatif pour les densités en-dehors de cet intervalle. Cette tendance ne saurait se révéler problématique que pour quelques variables dont la corrélation avec la densité de numéros résidentiels est très élevée.

### Biais obtenus en utilisant des poids tronqués

La troncation des poids peut entraîner l'introduction de biais importants, selon le rapport entre les variables à estimer et les poids qui sont tronqués. Dans certains cas, le biais dû à la troncation peut limiter l'ampleur de celle-ci et, par conséquent, restreindre son utilité comme méthode de réduction de la variance.

Nous avons également calculé le biais relatif dans la taille de l'échantillon pour des échantillons de 10, 30 et 60 numéros par grappe, en utilisant des poids tronqués à une valeur environ trois fois supérieure au poids moyen. Dans le cas des échantillons de dix numéros par grappe, la troncation a été faite selon un facteur de 2 plutôt que de 3, comme nous l'avons vu précédemment. La différence entre le biais relatif dans la taille de l'échantillon observé lorsqu'on utilise des poids tronqués et lorsqu'on se sert de poids non tronqués est sans importance pour toutes les tailles d'échantillons prélevés dans les grappes et pour la plupart des valeurs de  $N_i/100$ . La seule différence perceptible a été remarquée lorsque la densité de numéros résidentiels était inférieure à 10 à 15%. Les risques de biais sont légèrement plus élevés dans ces régions. Toutefois, le biais relatif dans la taille de l'échantillon observé lorsqu'on utilise les poids tronqués est encore bien inférieur à 1%, pour toutes les densités de numéros résidentiels.

## 5. CONCLUSION

La version originale de la méthode de Mitofsky et Waksberg constitue une méthode efficace pour obtenir par composition aléatoire un échantillon autopondéré de taille fixe. Toutefois, la nécessité de procéder à un contrôle séquentiel du nombre de ménages échantillonnés par grappe représente une caractéristique opérationnelle peu commode. En raison de ce contrôle, il est difficile de terminer la collecte des données dans des délais serrés. La période de collecte des données doit être suffisamment souple pour permettre l'obtention du nombre approprié d'unités dans chaque grappe. Le prolongement de la période de collecte des données et le contrôle du nombre d'unités échantillonnées se traduisent aussi par une augmentation des coûts. Etant donné la nécessité d'effectuer un contrôle fréquent du nombre d'unités échantillonnées, l'échantillonnage progressif peut également engendrer une certaine frustration chez les membres de l'équipe d'enquête en raison des complications relatives à la combinaison des opérations d'échantillonnage et de collecte des données.

La version modifiée de la méthode de Mitofsky et Waksberg élimine le caractère progressif du plan de sondage et, par le fait même, la nécessité de contrôler les travaux par grappe. Suivant cette méthode, on attribue un nombre fixe de numéros de téléphone à chaque grappe échantillonnée. Ainsi, on évite les dépenses afférentes au contrôle du nombre d'unités échantillonnées et à la mise en oeuvre d'une période de collecte des données prolongée. Toutefois, la version modifiée de la méthode de Mitofsky et Waksberg introduit de nouveaux éléments de biais et de variance dans les estimations. Il faut régler ces problèmes statistiques avant d'employer la méthode modifiée.

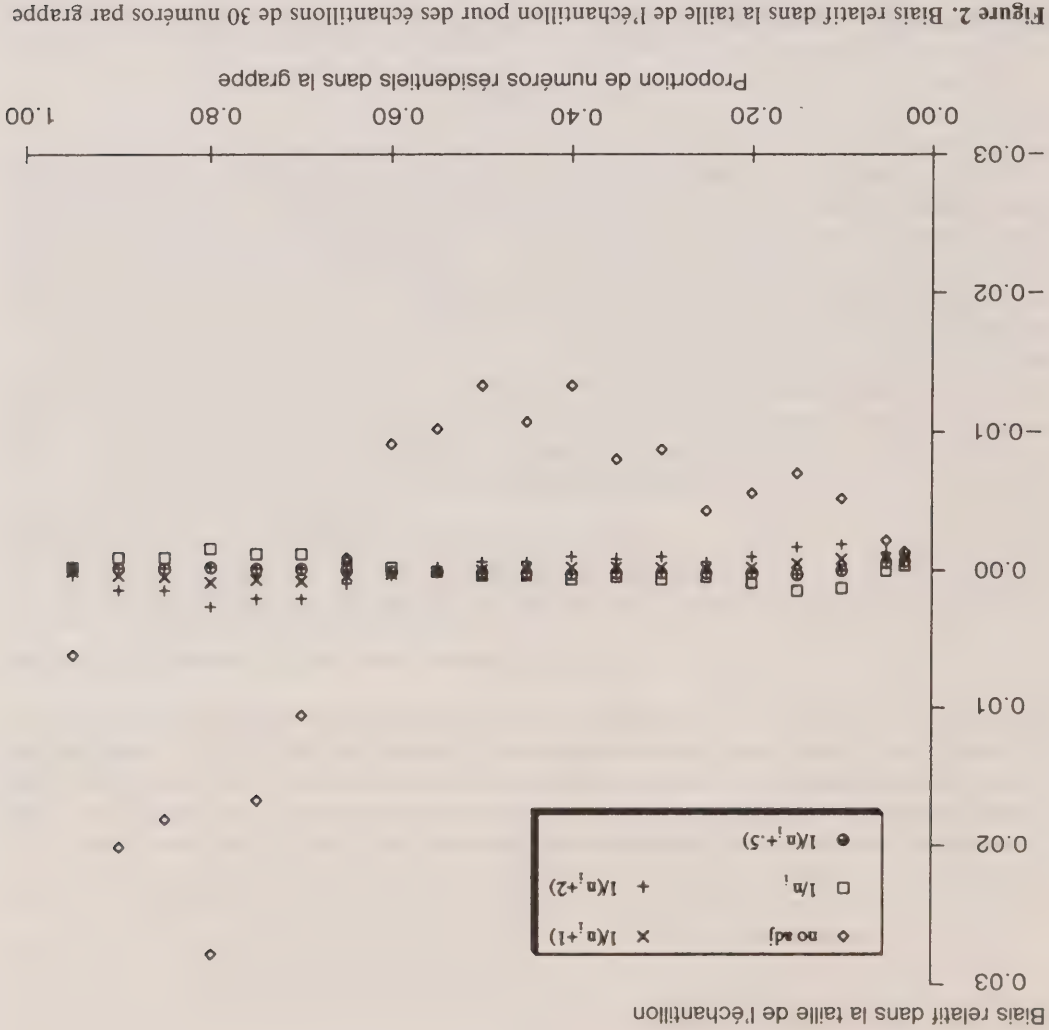


Figure 2. Biais relatif dans la taille de l'échantillon pour des échantillons de 30 numéros par grappe

à faible et à haute densité, les biais peuvent être assez élevés. Le tableau 1 de Cummings (1979) illustre, pour certaines variables, le biais introduit dans les estimations par suite de l'utilisation de poids non corrigés. Les biais ne sont pas très élevés, mais une pondération appropriée permettra de les éliminer.

En général, le rapport entre l'estimation et le nombre de ménages dans une grappe sera inconnu et incohérent pour toutes les variables à estimer. Par conséquent, il est raisonnable de choisir un estimateur pour lequel le biais relatif dans la taille de l'échantillon est peu élevé pour toutes les valeurs de  $N_i$ . Lorsque le biais relatif dans la taille de l'échantillon pour un ensemble d'estimateurs est peu élevé, le choix de ces derniers peut alors être dicté par la variance.

### Biais obtenus en utilisant des poids différents

Nous avons calculé le biais relatif dans la taille de l'échantillon en utilisant des estimateurs différents pour des échantillons de 5, 10, 30 et 60 numéros de téléphone par grappe. Dans le cas des échantillons de 30 et 60 numéros, le biais est négligeable sauf lorsqu'on utilise les poids non corrigés. On pourrait employer pour les échantillons de cette taille n'importe quel des estimateurs utilisant des poids corrigés sans introduire de biais dans les estimations. Dans le cas

L'histogramme ombré de la figure 1 illustre la distribution des ménages dont il est fait état dans le tableau 3. Nous l'avons superposé aux courbes des poids afin d'illustrer le fait que c'est dans les grappes qui ne contiennent qu'une très petite fraction des ménages échantillonnés que l'on observe les écarts les plus marqués entre les poids.

### Biais dans la taille de l'échantillon et dans les estimations

Dans presque toutes les enquêtes téléphoniques à composition aléatoire, y compris celles dans lesquelles on utilise le plan de sondage de Mitofsky et Waksberg, on procède à une stratification a posteriori de l'échantillon selon les totaux connus de personnes ou de ménages. L'un des principaux objets de cette mesure consiste à rectifier les estimations en fonction de l'ensemble de la population plutôt que de les faire porter uniquement sur les ménages ayant le téléphone. Massey et Botman traitent de cette question et exposent d'autres avantages de l'emploi de la stratification a posteriori dans les enquêtes téléphoniques à composition aléatoire au chapitre 9 de Groves et coll. (1988).

Peu importent les raisons de l'utiliser, la stratification a posteriori permet d'obtenir des estimations qui sont égales aux totaux connus, quels que soient les poids appliqués aux ménages individuels. Comme ce biais, qu'on peut considérer comme un biais dans la taille de l'échantillon, est toujours nul, il est difficile de trouver une seule statistique qui mesure directement le biais inconditionnel. Nous allons tenter de surmonter cette difficulté en examinant de quelle façon le biais dans la taille de l'échantillon varie en fonction de la densité de ménages dans la grappe.

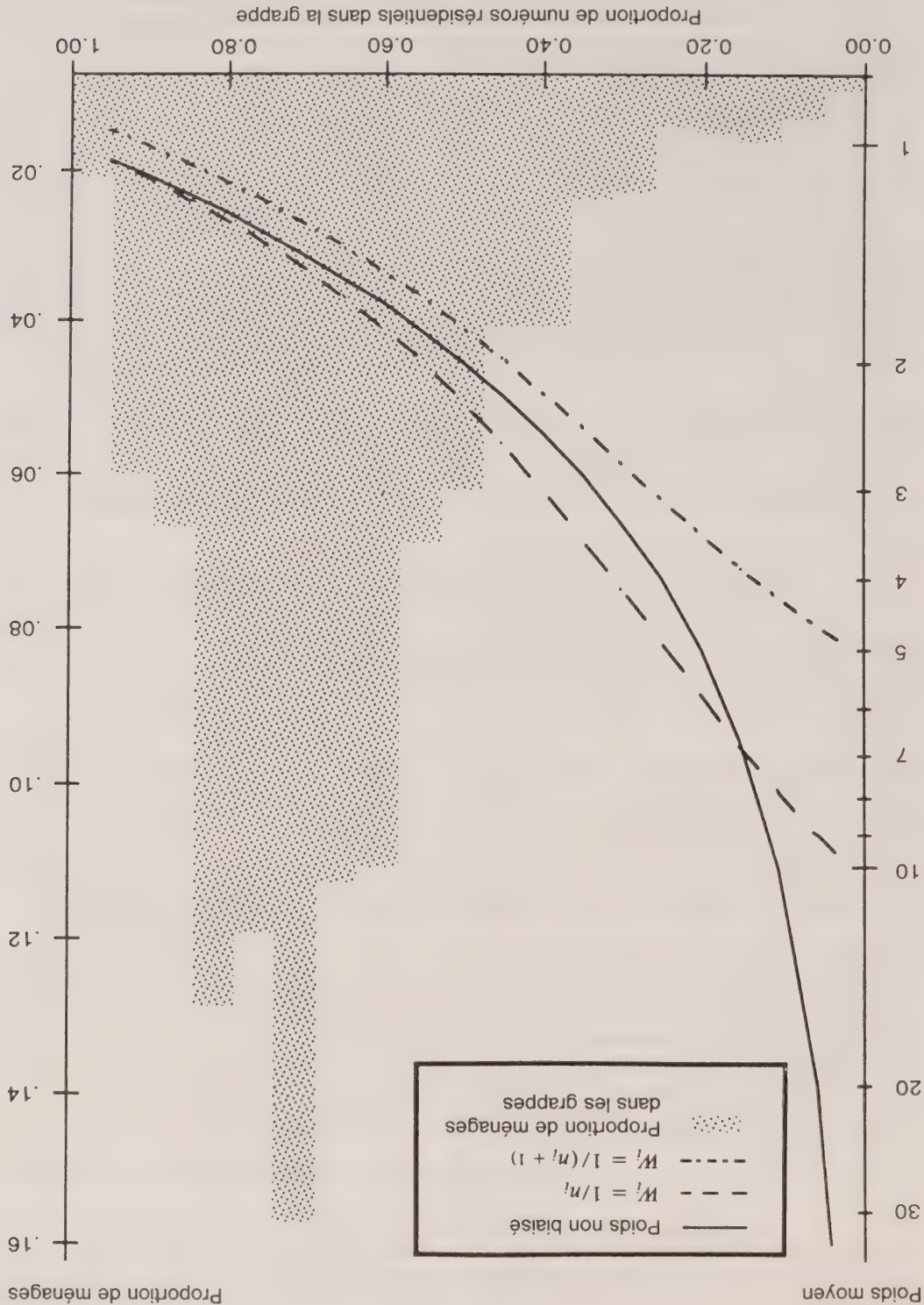
Nous avons évalué cette contribution au biais dans la taille de l'échantillon en suivant les étapes ci-après. D'abord, nous avons calculé les différents estimateurs ou fonctions de pondération à l'aide de la densité empirique de ménages indiquée au tableau 3. Ensuite, nous avons soumis les estimations à une stratification a posteriori afin de les rendre égales à l'unité et nous avons déterminé la contribution au total pour différentes valeurs de  $N_i/100$ . Enfin, nous avons défini le biais relatif dans la taille de l'échantillon comme étant la différence entre la contribution au total pour un estimateur donné et la contribution au total lorsqu'on utilise  $W_i$  comme poids. Cette mesure tient donc compte à la fois de la différence entre les poids pour des valeurs fixes de  $N_i$  et de la distribution des ménages pour toutes les valeurs de  $N_i$ . Par conséquent, les ménages échantillonnés dans des grappes ayant des valeurs de  $N_i$  qui sont rares ne contribueront pas beaucoup au biais relatif dans la taille de l'échantillon, même s'ils sont associés à de grands écarts entre les poids.

Pour illustrer ces calculs, la figure 2 montre le biais relatif dans la taille de l'échantillon pour certains estimateurs utilisés pour des échantillons de 30 ménages par grappe. L'un des estimateurs utilise le poids non corrigé, c'est-à-dire que le poids est constant pour tous les ménages, quel que soit le nombre de ménages relevés dans une grappe. Le biais relatif dans la taille de l'échantillon pour l'estimateur utilisant des poids non corrigés est beaucoup plus important que pour l'estimateur utilisant d'autres poids. Lorsqu'on utilise le poids non corrigé, les biais relatifs dans la taille de l'échantillon varient d'environ  $-2\%$  à  $+3\%$ .

L'importance du biais entachant l'estimation d'une caractéristique est limitée par l'amplitude du biais dans la taille de l'échantillon. Autrement dit, le biais relatif entachant l'estimation ne peut pas être plus élevé que celui dans la taille de l'échantillon. Cette limite supérieure est atteinte seulement lorsqu'il existe une corrélation parfaite entre la variable et la densité de numéros résidentiels, ce qui est très rare. Il est peu probable qu'il existe des corrélations très élevées dans les échantillons tirés à l'échelle nationale, mais il y a plus de chances d'en retrouver dans les échantillons prélevés dans des régions géographiques restreintes.

On peut constater à la figure 2 que certaines tendances se dégagent du graphique des biais: ainsi, l'estimateur utilisant des poids non corrigés est uniformément trop bas dans les grappes à faible proportion de numéros résidentiels et trop élevé, dans les grappes où la proportion de ménages est grande. Lorsque les valeurs prises par les variables varient entre les grappes





Malgré les lacunes opérationnelles que comporte l'augmentation de la taille de l'échantillon, nous l'avons quand même soumis à une examen limité. Comme elle n'a pas permis d'obtenir de meilleurs résultats que la troncation des poids, nous n'avons pas cru bon d'en faire une étude plus poussée.

#### 4. EFFETS DE L'UTILISATION DE LA VERSION MODIFIÉE DE LA MÉTHODE DE MITOFSKY ET WAKSBERG SUR LE BIAIS

L'augmentation de la variance n'est qu'une des conséquences de l'utilisation de la version modifiée de la méthode d'échantillonnage de Mitofsky et Waksberg. Cette méthode a aussi pour conséquence d'introduire un biais dans les estimations résultantes. Si on prélève un nombre fixe de numéros dans chaque grappe sans apporter aucune correction aux poids, la variance des estimations n'augmente pas, mais le biais risque d'être très élevé.

Le poids non biaisé ( $W_n$ ) pour la version modifiée de la méthode est défini par

$$W_n = \frac{100}{100} \times \frac{N_i}{K},$$

où les termes correspondent aux termes définis plus haut. La difficulté tient au fait que  $N_i$  est inconnue et ne s'annule pas avec le terme utilisé au deuxième degré, comme dans le cas de la version originale de la méthode de Mitofsky et Waksberg. Il faut donc avoir recours à la pondération pour tenter de réduire le biais.

Nous désignons l'estimateur qui utilise un poids de  $n_i^{-1}$  comme l'estimateur naturel parce que, pour un échantillonnage à partir d'une distribution binomiale ou hypergéométrique,  $n_i/K$  est un estimateur non biaisé de  $N_i/100$ . (Nous définissons le poids par  $n_i^{-1}$ , bien qu'il corresponde en fait à  $K/100n_i$ . Étant donné que  $K/100$  est une constante, l'utilisation de  $n_i^{-1}$  n'a aucun effet sur le rapport entre les poids.) Ce poids semble être l'estimateur naturel, malgré le fait que  $n_i^{-1}$  constitue une estimation biaisée de  $N_i^{-1}$ , à moins que les 100 numéros contenus dans une grappe soient tirés. Le biais introduit par  $n_i^{-1}$  a été étudié dans divers ouvrages; à cet égard, on peut se reporter à l'étude sur la stratification postérieure à l'échantillonnage dans Hansen, Hurwitz et Madow (1953). Il est peu probable qu'il existe un estimateur simple non biaisé, certainement pas de la forme  $(n_i + t)^{-1}$ , pour toutes les tailles d'échantillons prélevées dans les grappes.

Il est possible d'évaluer le biais susceptible d'être introduit en comparant la valeur prévue des estimateurs (le poids moyen calculé à l'aide d'estimateurs de la forme  $(n_i + t)^{-1}$  avec le poids non biaisé,  $W_n$ . Étant donné que tant le poids non biaisé que la valeur prévue des estimateurs sont des fonctions de  $N_i$ , nous étudierons d'abord ces quantités étant donné  $N_i$ . Le graphique de la figure 1 montre les courbes du poids non biaisé et des poids moyens obtenus en utilisant les estimateurs  $n_i^{-1}$  et  $(n_i + 1)^{-1}$ , lorsque  $K = 10$  numéros de téléphone sont sélectionnés dans chaque grappe. La fraction de sondage constante par grappe,  $r$ , a été omise de tous les poids. Étant donnée l'étendue de  $W_n$ , nous avons représenté la variation des poids moyens selon une échelle logarithmique.

Le graphique montre clairement qu'on enregistre les plus grands écarts entre  $W_n$  et les poids moyens obtenus à l'aide des deux estimateurs lorsque  $N_i/100$  est peu élevé. Dès que la densité de numéros résidentiels dépasse 20% (lorsqu'on utilise  $(n_i + 1)^{-1}$  et 10% (lorsqu'on utilise  $n_i^{-1}$ ), les écarts sont relativement mineurs. Bien que le graphique indique que le poids obtenu à l'aide de l'estimateur  $(n_i + 1)^{-1}$  est toujours inférieur à  $W_n$ , il en ira autrement si l'on procède à une stratification à posteriori. Les poids stratifiés à posteriori ne sont pas illustrés dans le graphique parce que la stratification à posteriori a réellement effet sur les poids conditionnels plutôt que sur les poids conditionnels illustrés ici. Nous aborderons la question du biais inconditionnel ci-après.

Facteurs d'inflation de la variance (FIV) approximatifs pour les échantillons obtenus par composition aléatoire suivant la version modifiée de la méthode de Mitofsky et Waksberg, avec des poids tronqués\*

Poids	Taille de l'échantillon prélevé dans chaque grappe (K)			
	10	30	60	100
$1/n_i$	1.12	1.11	1.09	1.09
$1/(n_i + .5)$	1.11	1.10	1.09	1.09
$1/(n_i + 1)$	1.09	1.10	1.09	1.09
$1/(n_i + 2)$	1.07	1.09	1.09	1.08

\* Tous les poids ont été tronqués à une valeur trois fois supérieure au poids moyen, sauf dans le cas des échantillons composés de dix numéros, où ils ont été tronqués à une valeur deux fois supérieure au poids moyen.

aléatoire menées par la société Westat, on a procédé à une troncation des poids à une valeur qui était de deux à trois fois supérieure au poids moyen. Pour les fins de la présente recherche, nous avons examiné les poids qui ont été tronqués à une valeur environ trois fois supérieure au poids moyen. Dans le cas des échantillons de dix numéros par grappe, les poids ont été tronqués à une valeur deux fois supérieure au poids moyen afin d'éviter que le nombre d'observations touchées soit trop faible.

Le tableau 5 indique les FIV des estimateurs pour différents nombres de numéros échantillon par grappe, lorsque le seuil de troncation est établi à une valeur trois fois supérieure au poids moyen pour  $n_i^{-1}$ . On n'y donne pas les FIV pour les échantillons composés de cinq numéros par grappe, parce que le seuil de troncation dans les échantillons de cette taille est presque égal à l'unité, le poids le plus élevé possible.

Les valeurs figurant dans le tableau montrent que la troncation permet de réduire considérablement les effets défavorables de la pondération différentielle sur la variance des estimations. C'est pour l'estimateur naturel qu'on observe la réduction la plus marquée du FIV, ce dernier diminuant de plus de 50% par suite de la troncation. Bien que les FIV des autres estimateurs diminuent quelque peu, ces baisses sont moins frappantes étant donné que ces FIV étaient déjà moins élevés que celui de l'estimateur naturel. La troncation peut toutefois introduire des biais susceptibles de neutraliser l'avantage relatif à la réduction de la variance. Nous aborderons cette question dans la section 4.

Variances obtenues par suite d'une augmentation de la taille de l'échantillon

Un troisième moyen de réduire la variabilité des poids consiste à augmenter la taille de l'échantillon. Les poids sont élevés lorsque le nombre de ménages prélevés dans la grappe est faible par rapport au nombre prévu de ménages par grappe. On peut faire diminuer la probabilité d'occurrence d'une telle situation en augmentant la taille de l'échantillon. Si le nombre de ménages dans une grappe est inférieur à une valeur déterminée (disons inférieur au tiers du nombre moyen de ménages par grappe), la taille de l'échantillon prélevé dans la grappe peut alors être doublée ou accrue dans une autre proportion.

On peut répéter cette opération afin de s'assurer que le nombre de ménages prélevés dans la grappe atteint une limite préalable ou jusqu'à ce que tous les numéros contenus dans la grappe soient utilisés. Ce plan itératif a pour inconvénient évident de nécessiter le contrôle du rendement de l'échantillon par grappe et de constituer un échantillonnage progressif. Cette méthode a aussi pour inconvénient d'entraîner l'échantillonnage d'un plus grand nombre de numéros de téléphone dans les grappes ayant une plus faible densité de ménages (celles où l'on risque le plus de devoir augmenter la taille de l'échantillon), et donc de provoquer une baisse de productivité.



Facteurs d'inflation de la variance (FIV) approximatifs pour les échantillons obtenus par composition aléatoire suivant la version modifiée de la méthode de Mitofsky et Waksberg

Tableau 4

Poids	Taille de l'échantillon prélevé dans chaque grappe (K)				
	100	60	30	10	5
$1/n_i$	1.17	1.23	1.29	1.34	1.31
$1/(n_i + .5)$	1.16	1.18	1.20	1.21	1.18
$1/(n_i + 1)$	1.14	1.15	1.16	1.15	1.12
$1/(n_i + 2)$	1.13	1.12	1.11	1.09	1.07

Lorsque la taille de l'échantillon prélevé dans chaque grappe est peu élevée, la moyenne de la variance conditionnelle est la composante dominante de la variance globale. À mesure que la taille de l'échantillon prélevé dans chaque grappe s'accroît, la variance de la moyenne conditionnelle devient plus dominante. C'est la raison pour laquelle la variance relative des poids, indiquée sur la première ligne du tableau 4, n'est pas une fonction monotone de la taille de l'échantillon prélevé dans chaque grappe.

Variances obtenues en utilisant des poids différents

Nous avons aussi examiné des poids autres que les poids inversement proportionnels au nombre de ménages afin de déterminer leurs effets sur l'erreur systématique et sur la variance des estimations. Bon nombre des poids de rééchantillonnage qui ont été étudiés ont été obtenus à partir des transformations de stabilisation de la variance proposées pour les variables binomiales. De toutes les solutions de rééchantillonnage qui ont été examinées, les estimateurs comportant le plus petit biais et la plus petite variance étaient obtenus en apportant de simples corrections au poids naturel. En particulier, l'addition d'une petite constante au nombre observé de ménages (estimateurs de la forme  $(n_i + t)^{-1}$ , où  $t$  est égal à .5, 1 ou 2) a permis de réduire les augmentations de la variance dues à la pondération différentielle. L'addition de la constante permet de réduire l'intervalle de variation des poids en tronquant les valeurs des poids les plus grands tout en ne modifiant que légèrement les poids des grappes où l'on trouve un plus grand nombre de ménages échantillonnés.

Le tableau 4 montre les FIV des estimateurs de la forme  $(n_i + t)^{-1}$  dans le cas où des nombres différents de numéros de téléphone sont échantillonnés dans chaque grappe. Les données contenues dans ce tableau sont aussi fondées sur les distributions des ménages et des grappes figurant au tableau 3. Ce tableau démontre clairement qu'il est possible de réduire considérablement la variance due à une pondération inégale en utilisant un estimateur de la forme  $(n_i + 1)^{-1}$  plutôt que l'estimateur naturel. Ceci est particulièrement vrai dans le cas des plans de sondage par composition aléatoire où l'on prélève 30 numéros de téléphone ou moins par grappe. En pareil cas, l'augmentation de la variance due à la pondération différentielle n'est que de 16% lorsqu'on utilise l'estimateur  $(n_i + 1)^{-1}$ , tandis qu'elle s'élève à 29% lorsqu'on utilise l'estimateur naturel.

Variances obtenues avec des poids tronqués

Un moyen souvent employé pour atténuer l'inflation de la variance associée à la variation des poids est la troncation des poids très élevés. Cette troncation, ou réduction des poids, est habituellement pratiquée à un poids au-dessus duquel on trouve relativement peu d'observations. Dans nombre d'échantillons prélevés au cours des enquêtes téléphoniques à composition

3. RÉPERCUSSIONS SUR LA VARIANCE DE L'UTILISATION  
DE LA VERSION MODIFIÉE DE LA MÉTHODE DE  
MITOFSKY ET WASKSBERG

Dans la version originale de la méthode de Mitofsky et Waksberg, la variance d'une estimation fondée sur l'échantillon est fonction du nombre de ménages tirés par grappe ainsi que de l'homogénéité des ménages à l'intérieur des grappes et entre celles-ci. La variance pour un échantillon aléatoire simple par  $[1 + \rho(n - 1)]$ , où  $\rho$  est le coefficient de corrélation interne intra-classe et  $n$  est le nombre moyen de ménages par grappe. Étant donné que les grappes de numéros de téléphone sont souvent reliées à des régions géographiques et tendent à être assez homogènes, le prélèvement d'un grand nombre de ménages par grappe peut s'avérer inefficace. Lorsqu'on utilise la version modifiée de la méthode de Mitofsky et Waksberg, on introduit une autre source de variance en permettant au nombre de ménages prélevés de varier d'une grappe à l'autre. Comme on l'a souligné à la section 2, le dénominateur de la fraction de sondage utilisée au deuxième degré ne s'annule pas avec le nombre de ménages dans la grappe (qui est proportionnel à la probabilité de sélection au premier degré) et les probabilités globales d'échantillonnage des ménages varient d'une grappe à l'autre.

En raison de la variabilité des taux globaux d'échantillonnage des ménages d'une grappe à l'autre, les variances des estimations sont plus grandes que lorsqu'on utilise la version originale de la méthode de Mitofsky et Waksberg, où la probabilité de tirage de chaque ménage est la même. Kish (1965) et Waksberg (1973) ont exposé des méthodes permettant d'évaluer l'augmentation de la variance d'une estimation causée par des probabilités de tirage inégales. On peut obtenir une approximation simple de la variance d'une estimation obtenue selon une méthode de pondération inégale (où les poids ne traduisent pas les fractions de sondage variables utilisées dans des strates choisies exprès pour réduire les variances d'échantillonnage) en multipliant la variance d'échantillonnage qu'on obtiendrait avec un échantillon autopondéré par un facteur d'inflation de la variance (FIV), défini par  $FIV = \{1 + \text{Varrel}(\text{poids})\}$ . Nous allons utiliser cette approximation ci-après pour examiner les répercussions sur la variance de l'emploi de la version modifiée de la méthode de Mitofsky et Waksberg.

La variance relative des poids a été calculée en deux étapes. Premièrement, la moyenne et la variance des poids ont été calculées étant donné un échantillonnage effectué à partir d'une distribution hypergéométrique tronquée (puisqu'il n'est pas possible d'obtenir un nombre nul de ménages si la grappe est échantillonnée au premier degré) définie par la densité de ménages dans la grappe et par la taille de l'échantillon prélevé dans la grappe. La moyenne et la variance inconditionnelles des poids ont ensuite été calculées pour l'ensemble de la distribution des ménages dans les grappes échantillonnées, qu'on trouve au tableau 3. La distribution des ménages dans l'échantillon est essentielle à l'évaluation du FIV.

Le poids naturel attribué à un ménage dans la version modifiée de la méthode de Mitofsky et Waksberg est proportionnel à  $n_i^{-1}$ , où  $n_i$  est le nombre de ménages observés dans la grappe-échantillon  $i$ . Ce poids peut varier selon des facteurs s'échelonnant d'autant que  $1/K$  jusqu'à 1, où  $K$  est le nombre de numéros de téléphone tirés dans une grappe. Le poids moyen est d'environ 1,5/ $K$ , puisque près de 65% des numéros qu'on retrouve dans les grappes échantillonnées sont des numéros résidentiels.

Si le nombre de numéros de téléphone échantillonnés par grappe se situe entre 5 et 30, l'augmentation de la variance due à la pondération est alors d'environ 30%. Le FIV diminue légèrement à mesure que le nombre de numéros échantillonnés par grappe s'accroît au-dessus de 30, pour atteindre environ 17% lorsque tous les numéros de la grappe sont échantillonnés. Le FIV ou la variance relative des poids est une fonction de la distribution du nombre de ménages entre les grappes et de la variabilité de l'échantillonnage aléatoire à l'intérieur des grappes. Nous avons rendu cette décomposition explicite en exprimant la variance des poids comme la somme de la variance conditionnelle des poids et de la variance de leur moyenne conditionnelle, où la variable de condition est la densité de ménages dans la grappe.



Tableau 3

Proportion de numéros résidentiels par grappe dans l'enquête de 1989  
(Selon un échantillon de 1,000 grappes, comportant chacune 30 numéros de téléphone)

Proportion de numéros résidentiels par grappe	Poids moyen de la grappe <sup>1</sup>	Répartition des ménages		Répartition des grappes	
		Pour- centage	Pourcentage cumulatif	Pour- centage	Pourcentage cumulatif

0	xx	0	0.0	0.0	0.5
.001 à .049	21.76 <sup>2</sup>	5	0.0	0.3	0.8
.05 à .099	8.70 <sup>2</sup>	18	0.1	0.6	1.4
.10 à .149	5.22 <sup>2</sup>	41	0.2	0.9	2.3
.15 à .199	3.73 <sup>2</sup>	48	0.2	0.8	3.1
.20 à .249	2.90	53	0.3	0.7	3.8
.25 à .299	2.37	144	0.7	1.6	5.4
.30 à .349	2.01	178	0.9	2.5	7.1
.35 à .399	1.74	408	2.1	4.6	10.5
.40 à .449	1.54	459	2.3	6.9	13.9
.45 à .499	1.37	840	4.3	11.2	19.5
.50 à .549	1.24	1,040	5.3	16.5	25.8
.55 à .599	1.14	1,926	9.8	26.3	36.5
.60 à .649	1.04	2,126	10.9	37.2	47.4
.65 à .699	0.97	3,255	16.6	53.8	62.9
.70 à .749	0.90	2,610	13.3	67.1	74.5
.75 à .799	0.84	3,022	15.4	82.6	87.1
.80 à .849	0.79	1,556	7.9	90.5	93.2
.85 à .899	0.75	1,458	7.4	98.0	98.6
.90 à .949	0.71	399	2.0	100.0	100.0
.95 à .999	xx	0	0.0	100.0	100.0
Total	xx	19,586	100.0	xx	xx
Taille moyenne des grappes <sup>3</sup>		19.68			
Effet du plan de sondage <sup>4</sup>		1.28			

<sup>1</sup> Le poids de la grappe correspond à la proportion moyenne de numéros résidentiels dans une grappe (c.-à-d. 0.653)  
<sup>2</sup> La troncation des poids ramènerait ceux-ci à 3.  
<sup>3</sup> La taille moyenne des grappes correspond à la taille moyenne des 995 grappes comprenant au moins un numéro résidentiel.  
<sup>4</sup> L'effet du plan de sondage est réduit à 1.12 lorsque le poids maximal est de 3.

Une autre particularité des pourcentages contenus dans les tableaux 1 à 3 est digne de mention. En effet, ces pourcentages reflètent la distribution selon la taille des grappes qui ont été sélectionnées dans l'échantillon, et non pas la distribution des grappes dans l'ensemble des États-Unis. L'utilisation d'un échantillonnage se faisant avec une probabilité proportionnelle à la taille entraîne un suréchantillonnage des grappes comportant une proportion élevée de numéros résidentiels et une sous-représentation des grappes ne comportant qu'un petit nombre de ces numéros. Il est possible de convertir une distribution qui représente l'échantillon en une distribution qui représente l'ensemble de la population en multipliant chaque pourcentage par les poids des grappes et en calculant la distribution en pourcentage des chiffres ainsi obtenus. Comme les poids sont exactement proportionnels à la réciproque du nombre de ménages par grappe qui ont répondu à toutes les questions, la conversion de la distribution des ménages en une distribution représentant l'ensemble de la population permet d'obtenir les pourcentages indiqués dans la distribution des grappes. La distribution des grappes dans l'échantillon est donc identique à celle des ménages dans la population.

Nous avons choisi de présenter tant les distributions de tous les ménages-échantillon que celles des ménages qui ont répondu à toutes les questions, parce qu'elles présentent toutes deux un intérêt pour les chercheurs. Les analyses dont il est fait état dans les sections 3 et 4 sont fondées sur les données du tableau 3.



Nombre de ménages par grappe qui ont répondu à toutes les questions de  
présélection dans l'enquête de 1989  
(Selon un échantillon de 1,000 grappes, comportant chacune 30 numéros de téléphone)

Nombre d'interviews par grappe	Poids moyen de la grappe <sup>1</sup>	Répartition des ménages		Répartition des grappes	
		Pour- centage	Pourcentage cumulatif	Pour- centage	Pourcentage cumulatif

Total	xx	15,030	xx	15,11	1,000	xx
0	xx	0	0	0.0	8	0.8
1 ou	7,572	6	0	0.0	3	0.3
3 ou 4	4,332	37	0.2	0.3	10	1.0
5 ou 6	2,75	126	0.8	1.1	22	2.2
7 ou 8	2,02	403	2.7	3.8	53	5.3
9 ou 10	1,59	688	4.6	8.4	72	7.2
11 ou 12	1,32	1,325	8.8	17.2	115	11.5
13 ou 14	1.12	1,987	13.2	30.4	147	14.7
15 ou 16	0.98	2,636	17.5	50.0	170	17.0
17 ou 18	0.85	2,692	17.9	65.9	154	15.4
19 ou 20	0.78	2,387	15.9	81.8	123	12.3
21 ou 22	0.70	1,673	11.1	92.9	78	7.8
23 ou 24	0.64	816	5.4	98.3	35	3.5
25 ou 26	0,55	254	1.7	100.0	10	1.0
27 ou 28	xx	0	0	100.0	0	0
29 ou 30	xx	0	0	100.0	0	0
Effet du plan de sondage <sup>4</sup>	xx	15,030	xx	15,11	1,000	xx

1 Le poids de la grappe correspond à la taille moyenne des grappes (c.-à-d. 15,115) divisée par le nombre de ménages qui ont répondu à toutes les questions de présélection dans la grappe *i*.  
2 La troncation des poids ramènerait ceux-ci à 3.  
3 La taille moyenne des grappes correspond à la taille moyenne des 992 grappes comprenant au moins un ménage qui a répondu à toutes les questions de présélection.  
4 L'effet du plan de sondage est réduit à 1,12 lorsque le poids maximal est de 3.

Il convient de souligner que le tableau 1 a été compilé à partir d'un échantillon de 15 numéros de téléphone par grappe, tandis que les tableaux 2 et 3 sont fondés sur un échantillon de 30 numéros de téléphone par grappe. À l'évidence, les estimations du pourcentage de numéros résidentiels dans une grappe composée de 15 numéros de téléphone seront sujettes à une erreur d'échantillonnage plus élevée que les estimations établies à partir de 30 numéros de téléphone. Toutefois, le nombre de grappes utilisées dans le tableau 1 est plus de deux fois supérieur au nombre utilisé dans les tableaux 2 et 3, ce qui devrait largement compenser l'effet de l'écart entre la taille des grappes.

Il existe deux différences entre les tableaux 2 et 3. D'une part, le tableau 2 montre la distribution des ménages qui ont répondu à toutes les questions de présélection (comme dans le cas du tableau 1), tandis que les données du tableau 3 sont fondées sur tous les ménages-échantillon. Le fait que seules les données des ménages ayant répondu à toutes les questions soient utilisées dans le tableau 2 se traduit par une réduction du nombre estimatif moyen de ménages par grappe et modifie toute la distribution. En outre, ce fait accroît la variabilité des estimations de la distribution puisque les chiffres reflètent les erreurs d'échantillonnage découlant à la fois de la distribution des ménages par grappe et de celle des taux de non-réponse par grappe. D'autre part, le tableau 2 (tout comme le tableau 1) indique le nombre de ménages par grappe, tandis que le tableau 3 indique le pourcentage de numéros résidentiels par grappe. Il était pratique d'exprimer le tableau 3 sous cette forme aux fins des analyses dont il est fait état ultérieurement dans le présent article.

**Tableau 1**  
Nombre de ménages par grappe qui ont répondu à toutes les questions de présélection par grappe dans l'enquête de 1986  
(Selon un échantillon de 2,427 grappes, comportant chacune 15 numéros de téléphone)

Nombre d'interviews effectués par grappe	Poids moyen de la grappe <sup>1</sup>	Répartition des ménages		Répartition des grappes	
		Pourcentage	Pourcentage cumulatif	Pourcentage	Pourcentage cumulatif

0	xx	0	0.0	62	2.6
1	7,932	54	0.3	54	2.2
2	3,972	106	0.6	53	2.2
3	2,64	258	1.4	86	3.5
4	1,98	440	2.3	110	4.5
5	1,59	810	4.3	162	6.7
6	1,32	1,290	6.9	215	8.9
7	1.13	1,960	10.5	280	11.5
8	0.99	2,656	14.2	332	13.7
9	0.88	2,862	15.3	318	13.1
10	0.79	2,990	15.9	299	12.3
11	0.72	2,717	14.5	247	10.2
12	0.66	1,548	8.3	129	5.3
13	0.61	780	4.2	60	2.5
14	0.57	210	1	15	0.6
15	0.53	75	0	5	0.2
Total	18,756	100.0	xx	2,427	100.0
Taille moyenne des grappes <sup>3</sup>	7.93				
Effet du plan de sondage <sup>4</sup>	1.31				

<sup>1</sup> Le poids de la grappe correspond à la taille moyenne des grappes (c.-à-d. 7,93) divisée par le nombre de ménages qui ont répondu à toutes les questions de présélection dans la grappe *i*.  
<sup>2</sup> La troncation des poids ramènerait ceux-ci à 3.  
<sup>3</sup> La taille moyenne des grappes correspond à la taille moyenne des 2,365 grappes comprenant au moins un ménage qui a répondu à toutes les questions de présélection.  
<sup>4</sup> L'effet du plan de sondage est réduit à 1.12 lorsque le poids maximal est de 3.

le premier degré d'échantillonnage), comportant chacune 15 numéros de téléphone, pour un total de 36,405 numéros. Au total, 18,756 ménages ont répondu à toutes les questions de présélection et on a enregistré 2,396 cas de refus, 1,727 cas de non-réponse pour d'autres raisons, et 13,526 cas où le numéro composé était un numéro non résidentiel ou hors service, où l'appel n'a pas obtenu de réponse et où il a été impossible de classer le cas dans une des catégories. L'analyse porte uniquement sur les 18,756 ménages qui ont répondu à toutes les questions. Les données qui figurent dans les tableaux 2 et 3 ont été recueillies après d'un échantillon, prélevé en 1989, qui se composait de 1,000 grappes comportant chacune 30 numéros de téléphone, pour un total de 30,000 numéros de téléphone, desquels 19,586 étaient des numéros résidentiels appartenant à des ménages ayant répondu à toutes les questions de présélection. Le tableau 2 montre la distribution des 15,030 ménages qui ont répondu à toutes les questions et le tableau 3, la distribution des 19,586 numéros résidentiels relevés dans les 1,000 grappes. Les poids attribués aux grappes sont exprimés par  $n/n_i$ ,  $n$  étant le nombre moyen de ménages par grappe. Il semble utile d'exprimer le poids sous cette forme, car ils indiquent les écarts observés par rapport à un échantillon autopondéré. Le seul effet du plan de sondage qui cause l'augmentation des variances est la variabilité des fractions d'échantillonnage. Les autres aspects du plan de sondage n'ont pas d'effet.



## 2. MÉTHODE DE RECHANGE POUR ÉTABLIR LA TAILLE DES GRAPPES AVEC LA TECHNIQUE DE MITOFSKY ET WASKBERG

Comme on l'a indiqué précédemment, la technique de Mitofsky et Waksberg exige qu'on prélève un nombre constant de numéros résidentiels par grappe (ou blocs de numéros) pour s'assurer d'obtenir un échantillon autopondéré. La solution de rechange proposée consiste à prélever dans chaque grappe un échantillon ( $K$ ) comportant un nombre constant de numéros de téléphone. Le premier degré d'échantillonnage reste inchangé et consiste toujours à prélever des grappes dont la probabilité de tirage est proportionnelle au nombre de ménages. Le nombre de numéros de téléphone par grappe étant constant, il est possible de désigner les numéros-échantillon à l'avance et d'éliminer ainsi le processus séquentiel. Il convient de noter qu'il est quand même nécessaire d'effectuer un suivi afin de déterminer quels numéros-échantillon sont des numéros résidentiels, et ce aux deux degrés d'échantillonnage. Toutefois, ce suivi ne s'applique qu'à un ensemble fixe de numéros de téléphone et ne nécessite pas la mise en oeuvre d'un processus séquentiel.

En revanche, la méthode de rechange ne permet pas d'obtenir un échantillon autopondéré. Comme, au premier degré, les grappes sont tirées avec une probabilité proportionnelle à la taille, la probabilité qu'une grappe soit sélectionnée est de  $r N_i / 100$ , où  $r$  est le taux de sondage utilisé pour le tirage des grappes, c'est-à-dire la fraction de sondage qui s'applique au premier degré, et  $N_i$  est le nombre de numéros résidentiels compris dans la grappe  $i$ . Le poids devrait être proportionnel à  $N_i^{-1}$ , mais, comme  $N_i$  n'est pas connu, on considère qu'il est proportionnel à  $n_i^{-1}$ , le nombre de ménages-échantillon dans la grappe.

Cette version modifiée de la méthode de Mitofsky et Waksberg présente des caractéristiques intéressantes du point de vue opérationnel. Elle est simple. Elle permet virtuellement de prélever l'échantillon d'avance sans qu'aucune opération de contrôle coûteuse ne soit nécessaire. Enfin, bien qu'elle exige qu'on effectue une pondération, on peut obtenir les poids directement à partir des données-échantillon et les produire mécaniquement sans contrôle systématique par des spécialistes.

Cette version présente cependant quelques graves inconvénients. Premièrement, on introduit un biais lorsqu'on estime  $N_i^{-1}$  à l'aide de  $K/100n_i$ , où  $K$  est le nombre de numéros de téléphone sélectionnés par grappe (un nombre constant dans toutes les grappes). Bien que ce biais soit relativement faible, il existe quand même. On ne peut ni l'éliminer ni l'atténuer en apportant des modifications mineures aux poids en utilisant, par exemple,  $1/(n_i + 1)$  au lieu de  $n_i^{-1}$ , où " $1$ " désigne une constante fixe. Deuxièmement, l'introduction de poids variables augmente considérablement les variances des estimations. (L'augmentation n'est pas tant causée par les poids mêmes que par le fait qu'ils expriment des probabilités d'échantillonnage variables). Enfin, en employant la version modifiée, on perd l'une des caractéristiques utiles de la méthode de Mitofsky et Waksberg: la possibilité d'établir la taille exacte de l'échantillon. En effet, comme la méthode prévoit le prélèvement d'un nombre constant de ménages par grappe, il suffit de prélever un échantillon comportant le nombre approprié de grappes pour obtenir un échantillon ayant la taille désirée. Dans le cas de la version modifiée, la taille de l'échantillon devient une variable aléatoire qui, en règle générale, ne sera pas exactement égale à la valeur désirée. Bien que les écarts soient habituellement petits, il est utile de pouvoir obtenir exactement les tailles visées lorsque, aux termes de contrats ou d'engagements budgétaires, l'organisation responsable de l'enquête doit satisfaire à certaines exigences en la matière. Nous aborderons ces questions aux sections 3 et 4.

Avant de passer à l'étude des variances et des biais, il est utile d'examiner la distribution des grappes selon la taille pour l'ensemble des États-Unis. Les tableaux 1 à 3 présentent les estimations de ces distributions qui ont été calculées à partir des données recueillies lors de deux grandes enquêtes téléphoniques nationales à composition aléatoire réalisées aux États-Unis par la société Westat. Lors de ces deux enquêtes, on a employé la version modifiée de la méthode de Mitofsky et Waksberg qui a été décrite précédemment. L'échantillon de l'enquête sur laquelle porte le tableau 1, qui a été prélevé en 1986, se composait de 2,427 grappes (retenues après



taille de l'échantillon. Ce processus est particulièrement incommode lorsque l'on ne dispose que d'un temps limité pour la collecte des données.

Plusieurs auteurs ont tenté de modifier la méthode de Mitofsky et Waksberg afin d'atténuer ou d'éliminer le caractère progressif de l'échantillonnage. Porthoff (1987) a élaboré une généralisation de la technique de Mitofsky et Waksberg selon laquelle on choisit  $c$  numéros de téléphone par grappe, plutôt qu'un seul selon la méthode de Mitofsky et Waksberg, pour déterminer si l'on doit retenir celle-ci ou non. On s'assure d'obtenir un échantillon autopondéré en soumettant les grappes dans lesquelles un seul des  $c$  numéros de téléphone a été composé à un plan de sondage prévoyant le tirage d'un nombre fixe de ménages par grappe, et en soumettant les autres grappes à un plan prévoyant le tirage d'un nombre fixe de numéros de téléphone par grappe. Il n'est pas nécessaire de soumettre le dernier groupe de grappes à un échantillonnage progressif. Selon Porthoff, en pratique, la plupart des grappes tomberont dans la seconde catégorie, de sorte que le nombre d'opérations séquentielles nécessaires s'en trouvera nettement réduit, sans toutefois que ces dernières soient complètement éliminées.

Lepkowski et Groves (1986) décrivent une méthode d'échantillonnage selon laquelle on tire des blocs de numéros de téléphone, dont un nombre assez important figurent dans les annuaires (et dans d'autres sources, s'il y a lieu), avec une probabilité de tirage proportionnelle au nombre de numéros inscrits. Les blocs de numéros qui ne contiennent aucun numéro de téléphone ou qui n'en contiennent que très peu sont échantillonnés à l'aide de la méthode de Mitofsky et Waksberg. Sudman (1973) avait précédemment proposé l'échantillonnage de blocs de numéros avec une probabilité de tirage proportionnelle au nombre de numéros inscrits dans les annuaires, mais il n'avait rien prévu en ce qui a trait aux blocs vides (qui pourraient contenir des numéros non inscrits). Drew et Jaworski (1986) décrivent une enquête téléphonique à composition aléatoire qui a été menée au Canada et dans laquelle on a utilisé des listes de numéros résidentiels (inscrits dans les annuaires ou non) obtenues à titre onéreux pour servir de mesures de l'effectif. Étant donné que les listes étaient considérées comme étant pratiquement complètes, il n'était pas nécessaire de prélever des blocs vides. Pour autant que nous le sachions, il est impossible d'obtenir des dénombrements pratiquement complets des numéros résidentiels aux États-Unis. Ni le plan de sondage de Porthoff ni celui de Lepkowski et Groves n'éliminent entièrement la nécessité d'un échantillonnage progressif, bien que tous deux semblent réduire la part de l'échantillon devant faire l'objet d'un tel échantillonnage. Ces deux méthodes présentent aussi d'autres inconvénients. La technique de Porthoff semble plutôt complexe: pour autant que nous le sachions, elle n'a pas été beaucoup utilisée pour les enquêtes téléphoniques à composition aléatoire. Dans le cas des enquêtes nationales, la technique de Lepkowski et Groves nécessite, aux fins de l'obtention de mesures de l'effectif, l'achat et le dépouillement d'une liste-annuaire englobant tous les États-Unis. Bien qu'on puisse se procurer de telles listes commerciales, elles sont coûteuses. En outre, selon un certain nombre de rapports récents, le pourcentage de l'ensemble des numéros résidentiels qui sont inscrits dans les annuaires n'est pas très élevé, et il diminue rapidement. Tucker (1989) fait état d'une analyse des numéros inscrits dans un groupe de comtés et de villes des États-Unis, analyse qui révèle des taux d'inscription variant de 48 à 62%. Selon Linda Piekariski (1990), si le nombre de numéros non inscrits continue d'augmenter au rythme actuel, il se pourrait que, «d'ici l'an 2000, la proportion de ménages dont le numéro de téléphone ne sera pas inscrit s'élève jusqu'à 62%» (traduction). Ainsi, les mesures de l'effectif ne correspondent probablement que plus ou moins au nombre réel de ménages dans un bloc de numéros actifs.

Waksberg a proposé d'apporter à la technique de Mitofsky et Waksberg, une autre modification qui élimine complètement la nécessité de procéder à un échantillonnage progressif. La société Westat a eu recours à cette méthode dans un grand nombre d'études réalisées à l'aide de la composition aléatoire. Cummings (1979) avait déjà utilisé les mêmes techniques à la suite d'une erreur d'application de la méthode de Mitofsky et Waksberg. Ce dernier n'a toutefois pas reconnu que la méthode pouvait être utile pour éliminer l'échantillonnage progressif, ni examiné ses caractéristiques en vue de l'utiliser dans d'autres enquêtes. Nous allons maintenant décrire cette méthode et en exposer les propriétés mathématiques et statistiques.

# Méthode pour éviter l'échantillonnage progressif dans une enquête téléphonique à composition aléatoire

J. MICHAEL BRICK et JOSEPH WAKSBERG<sup>1</sup>

## RÉSUMÉ

La technique de Mitofsky et Waksberg est une méthode efficace pour prélever par composition aléatoire un échantillon autopondéré de ménages. Il s'agit d'une méthode progressive qui exige le prélèvement d'un nombre constant de ménages dans chaque grappe. Dans le présent article, les auteurs décrivent une version modifiée de la méthode de Mitofsky et Waksberg qui n'est ni autopondérée, ni progressive. Après avoir analysé les biais et la variance des estimations obtenues à l'aide de la méthode modifiée, ils indiquent dans quelles circonstances il peut être avantageux d'utiliser la version modifiée plutôt que la version originale de la méthode de Mitofsky et Waksberg.

MOTS CLÉS: Composition aléatoire; échantillonnage téléphonique; échantillonnage par grappes; troncation.

## 1. INTRODUCTION

La méthode élaborée par Mitofsky et Waksberg pour tirer des échantillons de ménages par composition aléatoire (Waksberg 1978) est souvent utilisée pour prélever un échantillon dans les enquêtes téléphoniques. Comme Waksberg le dit dans son article, il s'agit d'une méthode efficace de production d'un échantillon autopondéré, c'est-à-dire un échantillon dans lequel tous les ménages ayant le téléphone ont la même probabilité de tirage (à l'exception des ménages ayant plus d'un numéro de téléphone). L'efficacité de la méthode de Mitofsky et Waksberg est attribuable au fait qu'elle permet de réduire de façon marquée la proportion de numéros de téléphone non-résidentiels devant être composés avant qu'il soit possible de joindre les ménages-échantillon. La méthode de Mitofsky et Waksberg suit un plan de sondage à deux degrés. Au premier degré, on prélève un échantillon de grappes consistant chacune en un bloc de 100 numéros de téléphone ou en un multiple de tels blocs. Les grappes (ou blocs de 100 numéros de téléphone) sont d'abord tirées avec une probabilité égale. On choisit au hasard un numéro de téléphone dans chaque grappe et on le compose. S'il s'agit du numéro d'un ménage, on retient la grappe. Sinon, on la rejette. Au deuxième degré, on prélève un échantillon de ménages à l'intérieur des grappes-échantillon que l'on a retenues. Pour s'assurer d'obtenir un échantillon autopondéré, il faut prélever un nombre constant de ménages par grappe. Certaines organisations (dont Westat Inc.) vont en général un peu plus loin et précisent un nombre constant de ménages interviewés par grappe (ou de ménages soumis à une sélection préliminaire, si la première partie du processus de collecte des données consiste en une présélection). On estime que le fait de substituer un autre ménage tiré au hasard à l'intérieur de la même grappe à chaque ménage non répondant constitue une façon raisonnable de réduire le biais dû à la non-réponse.

Ce système présente un inconvénient sur le plan opérationnel. Il faut parfois recomposer un numéro un assez grand nombre de fois pour déterminer s'il s'agit d'un numéro résidentiel ou non, surtout dans le cas des numéros auxquels on n'obtient pas de réponse. Le nombre de rappels nécessaires pour déterminer quels ménages acceptent de collaborer est encore plus élevé. On doit procéder ainsi pour chaque échantillon initial afin de déterminer dans quelles grappes il faut ajouter des numéros de téléphone pour obtenir la taille voulue et quel nombre de numéros on doit ajouter. De fait, il faut utiliser dans chaque grappe un échantillonnage progressif selon lequel on doit régler tous les cas précédents avant de savoir s'il est nécessaire d'augmenter la

<sup>1</sup> J. Michael Brick et Joseph Waksberg, Westat Inc., 1650 Research Boulevard, Rockville, Maryland 20850, U.S.A.



l'emploi, la rémunération et les heures de travail, tandis que les laboratoires Bell procèdent actuellement à la mise au point d'un système IDTC pour une enquête auprès des clients de la société AT & T (Wendler 1990).

Le BLS utilise également la technologie de la reconnaissance de la parole à titre expérimental aux fins de la collecte des données de la CESS (voir Winter et Clayton 1990). Bien que les postes téléphoniques à clavier soient de plus en plus accessibles, nous estimons qu'entre 10% et 20% de nos répondants ne disposent pas de tels appareils. Dès que la technologie de la reconnaissance de la parole multilocuteur aura atteint un niveau acceptable pour les dix chiffres dont on a besoin pour déclarer les données de la CESS, nous nous attendons à ce que les utilisateurs préfèrent l'employer à la place du système de collecte par IDTC. Il convient donc d'entreprendre d'autres travaux de recherche sur les erreurs de mesure associées à cette technologie.

REMERCIEMENTS

Les auteurs remercient Darrell Philpot et Henry Chiang de leur aide dans cette étude.

BIBLIOGRAPHIE

CLAYTON, R., et HARRELL, L.J., Jr. (1990). Developing a cost model for alternative data collection methods: Mail, CATI and TDE. Présenté à l'Annual Meeting of the American Statistical Association, Anaheim, California.

COX, A.C., et COOPE, M.B. (1981). Selecting a voice for a specified task: The example of telephone announcements. *Language and Speech*, 24, 233-243.

MARICS, M.A., et WILLIGES, B.H. (1988). The intelligibility of synthesized speech in data inquiry systems. *Human Factors*, 30, 719-732.

PONIKOWSKI, C.H., COPELAND, K.R., et MEILY, S.A. (1989). Applications for touch-tone recognition technology in establishment surveys. Présenté à l'American Statistical Association Winter Conference, San Diego, California.

SCHWAB, E.C., NUSBAUM, H.C., et PISONI, D.B. (1985). Some effects of training on the perception of synthetic speech. *Human Factors*, 27, 395-408.

WATERWORTH, J.A. (1984). Interaction with machines by voice: A telecommunications perspective. *Behaviour and Information Technology*, 3, 163-177.

WENDLER, E.R. (1990). Respondent-initiated computer-directed surveys. Présenté à l'Annual Conference of the American Association of Public Opinion Research, Lancaster, Pennsylvania.

WERKING, G.S., TUPEK, A.R., PONIKOWSKI, C., et ROSEN, R. (1986). A CATI feasibility study for a monthly establishment survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 639-643.

WERKING, G., TUPEK, A., et CLAYTON, R. (1988). CATI and touchtone self-response applications for establishment surveys. *Journal of Official Statistics*, 4, 349-362.

WINTER, D.L.S., et CLAYTON, R.L. (1990). Speech data entry: Results of the first test of voice recognition for data collection. Bureau of Labor Statistics, Washington, D.C.



puisque'ils ont déclaré qu'ils étaient dérangés lorsqu'ils devaient répondre à l'enquête. Toutefois, étant donné que la plupart des répondants jugent le débit assez bon, et que bon nombre d'entre eux se servent de la fonction d'accélération, le fait de donner plus de temps pour répondre pourrait être une source de frustration. On ne peut probablement pas faire grand chose pour réduire le nombre de réintroductions, puisque les répondants indiquent qu'ils avaient appuyé sur la mauvaise touche et qu'ils devaient corriger les données.

Les données indiquent que le nombre d'erreurs diminue à mesure que les répondants s'habituent au système. Les enquêtes par panel sont donc peut-être celles qui se prêtent le mieux à ce mode de collecte des données. Par ailleurs, nous croyons que le système IDTC présente beaucoup de promesses pour les enquêtes auxquelles il faut répondre par oui ou non ou par des chiffres. Les erreurs relevées ne sont pas très graves, et les répondants ont évalué leur expérience de l'utilisation du système comme étant très favorable. La méthode IDTC peut être particulièrement attrayante pour les entreprises participant à des enquêtes, car elle leur permet d'appeler au moment qui leur convient, plutôt que d'être dérangées par des appels téléphoniques leur demandant de répondre à des questions. Toutefois, pour certaines enquêtes, l'auto-initiation et l'absence de contacts humains peuvent constituer des problèmes qui contribueraient à l'erreur due à la non-réponse.

Même si les répondants ont réservé un accueil très favorable à la collecte des données par IDTC, certaines mesures peuvent être prises afin d'améliorer le système, notamment:

- accorder suffisamment de temps aux répondants pour introduire leurs données au clavier, surtout dans le cas des premières questions et de celles nécessitant l'inscription d'une longue série de chiffres;
- chercher des moyens d'améliorer le mode de confirmation des réponses; enfin,
- prévoir des vérifications de contrôle longitudinales pour détecter les cas de déclaration de cents et d'autres grossières erreurs; on pourrait intégrer ces contrôles au système IDTC en posant aux répondants des questions appropriées afin de corriger ou de confirmer leurs réponses.

Le BLS a eu recours à la méthode IDTC dans une autre enquête, consistant en un suivi auprès d'un petit échantillon d'entreprises ayant participé, en 1988, à une enquête sur les programmes d'aide en matière de drogues mis en oeuvre par les employeurs. Réalisée en 1990, cette enquête d'aide avait pour objet de déterminer si le pourcentage d'entreprises qui offrent à leurs employés un programme d'aide en matière de drogues avait beaucoup changé au cours des deux années précédentes. Ces entreprises ont reçu par la poste un court questionnaire auquel elles devaient répondre par des données numériques ou par oui ou non, et on les a encouragées à déclarer leurs réponses au moyen d'un téléphone à clavier. À la fin des premières semaines de l'enquête, près de 20% des entreprises avaient répondu à l'aide du système IDTC, et une proportion égale par la poste. Le système IDTC n'a pas été utilisé après le début des activités de suivi des cas de non-réponse, soit environ deux semaines après l'envoi initial des questionnaires par la poste. Les autres données ont été recueillies par téléphone (méthode ITAO).

Compte tenu du temps que l'IDTC permet de sauver et des économies qu'elle permet de réaliser au titre des activités de perforation, nous croyons que la collecte des données au moyen d'un téléphone à clavier pourrait être utilisée avec profit pour d'autres enquêtes où la collecte des données et assujettie à des contraintes de temps. Il est possible que l'utilisation de ce mode de collecte communiquant au participant à l'enquête l'importance qu'il y a à respecter les délais de réponse. Nous avons démontré dans le présent article qu'on peut réduire les erreurs de mesure en utilisant le mode de collecte des données par IDTC.

Comme la collecte par IDTC permet de réduire les délais de réponses et les coûts de collecte, nous nous attendons à ce qu'elle soit beaucoup plus utilisée dans l'avenir. Nous savons que deux projets sont en cours afin de mettre à l'essai la reconnaissance des données introduites au moyen d'un téléphone à clavier dans le cadre d'une enquête. Statistique Canada procède actuellement à l'essai d'un système de collecte des données par IDTC pour l'enquête sur

Quant à la difficulté d'exécution de la tâche, 60% des répondants ont indiqué qu'ils n'avaient jamais eu à réintroduire de chiffres, tandis que la plupart des autres ont fait savoir qu'ils avaient parfois dû le faire. Lorsqu'on leur a demandé les raisons pour lesquelles ils devaient réintroduire des chiffres, une majorité ont répondu qu'ils avaient accidentellement appuyé sur une mauvaise touche. D'autres ont indiqué qu'ils n'avaient pas eu assez de temps, qu'ils avaient été distraits ou qu'ils avaient inscrit leurs chiffres trop rapidement. Dans les derniers mois de l'enquête d'évaluation, nous avons interrogé les répondants concernant la répétition des questions (sans qu'il n'y ait eu réintroduction des données). Environ 83% d'entre eux ont indiqué qu'ils n'avaient jamais jugé nécessaire de faire répéter les questions. La majorité des 17% de répondants qui ont fait répéter les questions ont déclaré qu'ils avaient été distraits, tandis que d'autres ont mentionné qu'ils avaient manqué de temps.

La plupart des répondants ont éprouvé peu de problèmes avec le système de télécommunications, puisque 93% d'entre eux ont indiqué que la communication téléphonique avait été bonne lorsqu'ils avaient utilisé le système IDTC. La plupart des répondants qui ont déclaré avoir éprouvé des problèmes de communication ont mentionné que cela ne s'était produit qu'une seule fois. Un grand nombre des répondants (63%) ont utilisé le symbole de la livre, fonction du système conçue pour accélérer la déclaration des données.

Presque tous les répondants ont indiqué que les instructions qu'ils avaient reçues lorsqu'ils ont commencé à utiliser le système IDTC étaient à-propos. En général, les répondants ont semblé satisfaits du système, environ 93% d'entre eux ayant évalué leur expérience avec le système IDTC comme étant très favorable.

## 6. ANALYSE

Selon les données recueillies, le mode de collecte par IDTC pose peu de problèmes graves. Les données de vérification des enregistrements indiquent l'existence d'une certaine erreur due à la non-réponse aux questions, erreur surtout associée aux répondants qui utilisent le système pour la première fois. L'inscription de chiffres additionnels ou inexacts semble être le plus grave problème qui touche les éléments d'information. Toutefois, dans une enquête par panel, les vérifications de contrôle longitudinales pourraient permettre de réduire le nombre de ces erreurs, comme pourraient le permettre les vérifications de contrôle logiques dans toutes les enquêtes. En outre, il est essentiel d'aborder la question de l'arrondissement des données dans les instructions transmises aux répondants. Tant les données de vérification des enregistrements que les données enregistrées mécanographiquement indiquent que les séries de chiffres plus longues posent plus de problèmes, probablement autant à l'étape de l'introduction des données qu'à celle de la vérification des données inexacts. Les répondants éprouvent peut-être de la difficulté à se rappeler des longues séries de chiffres durant la vérification, puisqu'ils ne semblent pas avoir de problèmes à comprendre les chiffres qui leur sont relus pour fin de confirmation.

D'après les données de vérification des enregistrements, il se peut que les répondants aient introduit le numéro de confirmation du mois dans la zone relative à la question sur le nombre total d'employés. En outre, les données enregistrées mécanographiquement indiquent qu'il arrive souvent que les répondants n'inscrivent pas le mois la première fois que celui-ci est demandé. Comme les répondants semblent se servir de leurs formules d'enquête lorsqu'ils introduisent des données, il est probable qu'ils aient besoin d'un peu plus de temps pour passer du numéro d'identification au haut de la formule, au mois et aux éléments d'information qui se trouvent plus bas. On pourrait résoudre ce problème en mettant tous les renseignements devant être introduits à un seul endroit sur la formule. Ainsi, on pourrait diminuer le nombre de fois que la question du "mois" doit être répétée et, potentiellement, réaliser des économies en réduisant la durée des appels. On pourrait réduire le nombre de fois que les autres questions doivent être répétées en donnant aux participants à l'enquête plus de temps pour répondre.



Quelques répondants ont appelé trois fois pour introduire leurs données pour un mois donné, et un répondant a même appelé quatre fois. Ces répondants semblaient éprouver des difficultés avec le système, mais ils ont quand même réussi à déclarer toutes leurs données correctement. En général, ces données laissent supposer que les répondants éprouvent certaines difficultés à utiliser le système (plus qu'ils ne l'admettent durant l'interview d'évaluation). Certaines mesures pourraient être prises pour aider à atténuer une partie de ces problèmes. On pourrait, entre autres, donner plus de temps pour répondre, fournir de meilleures instructions et essayer d'améliorer la méthode de confirmation des données introduites. En outre, il se peut qu'il suffise, pour résoudre quelques-uns des problèmes, de donner aux répondants la possibilité de revenir à une question.

### 5.3 Enquête d'évaluation auprès des répondants

Au cours de l'année 1990, les intervieweurs du BLS ont mené une enquête d'évaluation téléphonique auprès des répondants qui ont utilisé le système IDTC. Compte tenu des ouvrages sur l'ergonomie que nous avons étudiés précédemment, certaines des questions portaient sur la compréhension et le débit de la voix numérisée. Comme les données enregistrées mécano-graphiquement ont révélé un nombre substantiel de répétitions et de réintroductions, certaines questions ont aussi été élaborées afin d'étudier ce problème. En outre, on a demandé aux répondants d'évaluer le système IDTC et de répondre à des questions portant sur la conception du système.

Les résultats de l'enquête d'évaluation sont présentés dans le tableau 4. Les répondants n'ont eu que peu de difficulté à comprendre la voix numérisée. Environ 97% d'entre eux ont déclaré que la voix était très compréhensible, et ils ont tous indiqué qu'il était facile de comprendre les chiffres à mesure qu'ils leur étaient relus pour fin de confirmation. Au cours des deux premiers mois de l'enquête, nous avons interrogé les répondants au sujet du débit de l'interview. La plupart des répondants ont indiqué que le débit était à peu près correct (88%), bien que 10% d'entre eux aient estimé qu'il était trop lent. Tout au long de l'étude, nous avons supposé que la compréhension de la voix constituait un problème moins important que les difficultés à exécuter la tâche. Bien qu'il ait été difficile de séparer les deux types de problèmes dans les données de vérification des enregistrements, les interviewers d'évaluation sont venues confirmer cette hypothèse.

Tableau 4

Résultats de l'enquête d'évaluation\*

Voix compréhensible	97%
Chiffres relus faciles à comprendre	100%
Débit assez bon	88%
Jamais réintroduit de chiffres	60%
Jamais fait répétée de questions	83%
Jamais eu une mauvaise communication téléphonique	93%
Utilise la fonction d'accélération	63%
Instructions adéquates	98%
Expérience avec l'IDTC très favorable	93%

\* N = 411, à l'exception des rubriques relatives au débit et aux questions répétées. Environ 177 répondants ont été interrogés au sujet du débit et 209 concernant la répétition des questions.



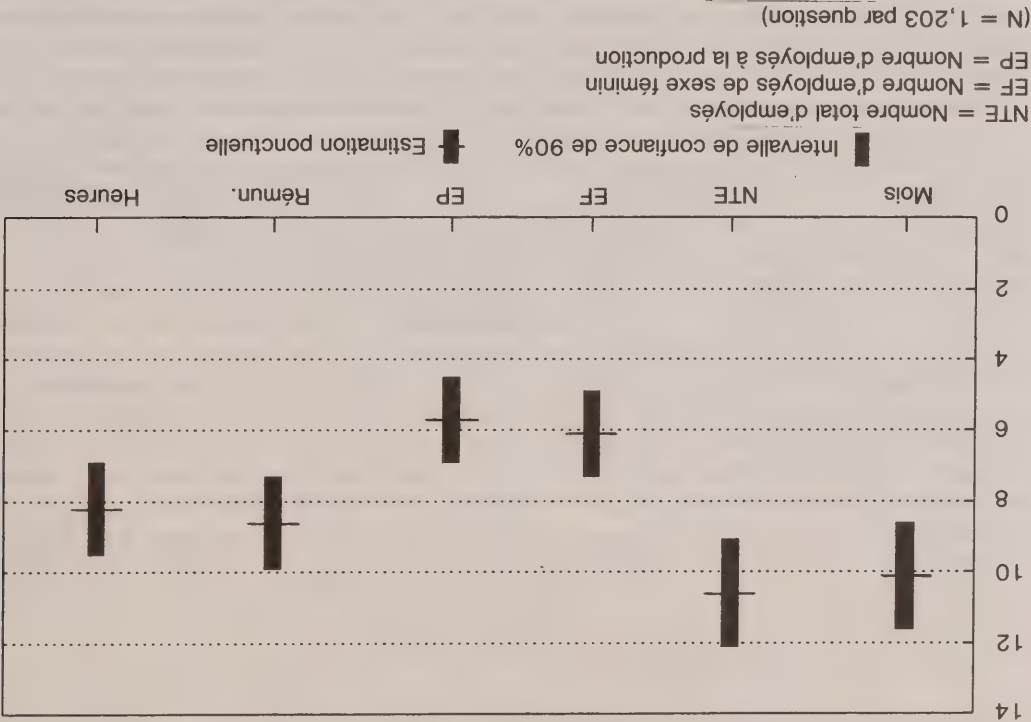


Figure 1. Données enregistrées mécanographiquement – Pourcentage de questions répétées et de réintroduction des données

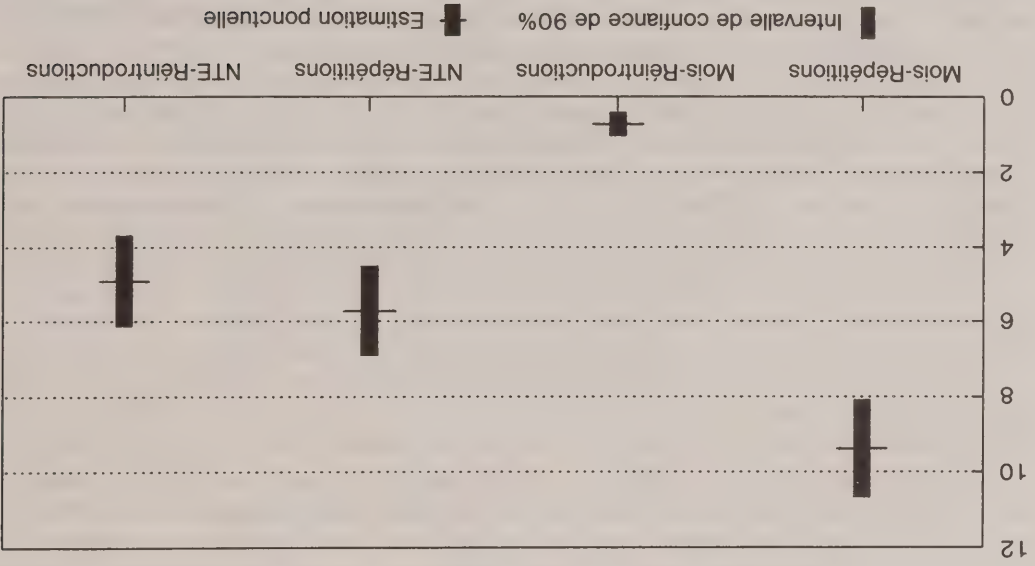


Figure 2. Données enregistrées mécanographiquement – Pourcentage de questions répétées par rapport aux réintroductions de données pour les deux premiers éléments d'information

deux ou trois mois. Comme les données enregistrées mécanographiquement variaient peu d'un mois à l'autre, toutes les données sont présentées pour les trois mois combinés.

La figure 1 indique, pour chaque question, le pourcentage d'appels pour lesquels la question a été posée au répondant plus d'une fois. La question pouvait être répétée une deuxième fois si le répondant ne répondait pas dans les deux secondes (répétition), ou s'il ne confirmait pas sa réponse en appuyant sur le "1" après que celle-ci lui eut été relue (réintroduction). La figure indique que les taux de répétitions et de réintroductions sont plus élevés dans le cas des questions portant sur le mois, sur le nombre total d'employés, sur la rémunération et sur les heures de travail. Il est possible que les taux plus élevés enregistrés pour les deux premières questions (plus de 10%) soient attribuables au fait que les répondants aient eu besoin de quelques questions pour se familiariser avec le système. Les questions sur la rémunération et sur les heures de travail nécessitant en général l'introduction d'un plus grand nombre de chiffres, nous supposons que les risques d'erreurs d'introduction des données sont plus élevés, ce qui explique pourquoi la question doit être répétée et la réponse réintroduite.

La figure 2 donne séparément les pourcentages de questions répétées faute de réponse (répétitions) et à défaut de confirmation (réintroductions) pour les deux premières questions ("mois" et "nombre total d'employés"). Il n'a pas été possible d'obtenir des données distinctes pour les autres éléments d'information. Lorsqu'ils utilisent le système IDTC, les répondants à la CESS sont tenus d'introduire au moins leur numéro d'identification de déclarant, le mois et le nombre total d'employés. Le système accepte une non-réponse aux autres questions. Le fait que les répondants doivent au moins indiquer le mois et le nombre total d'employés explique pourquoi nous avons pu obtenir des données distinctes sur les répétitions et sur les réintroductions dans le cas de ces deux questions. Presque tous les problèmes que posait la question relative au mois étaient dus à des répétitions, c'est-à-dire au fait que deux secondes se soient écoulées sans qu'une réponse soit donnée. Par ailleurs, on comptait parmi les problèmes posés par la question relative au nombre total d'employés presque autant de répétitions faute de réponse que de réintroductions à défaut de confirmation de la réponse précédente.

Seulement 2% des appels reçus par le système IDTC chaque mois comprenaient uniquement un numéro d'identification du déclarant. Lors de ces appels, les répondants avaient tout simplement raccroché ou avaient pu voir leur communication avec le système coupée.

En outre, le système IDTC enregistre tous les appels reçus qui comprennent au moins le numéro d'identification du déclarant, le mois et le nombre total d'employés. À partir des données de vérification des enregistrements dont il a été question précédemment, nous avons pu repérer les répondants qui ont effectué plus d'un appel durant un mois et codé les raisons des rappels. En tout, environ 4% des répondants ont appelé le système plus d'une fois au cours d'un mois donné. La plupart de ces répondants ont fourni des éléments d'information qu'ils n'avaient pas communiqués lors de l'appel initial (2%). Une proportion additionnelle de 1% des répondants ont corrigé certains éléments d'information en plus d'en fournir de nouveaux. Bon nombre de ces répondants ont semblé avoir éprouvé des problèmes à introduire les données la première fois. Une dernière proportion de 1% des répondants n'ont rappelé que pour apporter des corrections aux éléments d'information fournis antérieurement ou pour transmettre des données identiques. Ces appels étaient souvent effectués plusieurs jours plus tard, ce qui veut peut-être dire que les répondants avaient obtenu de nouvelles données de leurs dossiers. Pour ce qui est des données identiques, les répondants ne se souvenaient peut-être pas s'ils avaient déjà déclaré leurs données ou non. Actuellement, le système accepte les données transmises à la date et à l'heure les plus récentes, même si les analystes obtiennent la liste des répondants qui ont rappelé à des fins de révision et, au besoin, de correction.

Il semble que la nécessité d'introduire un "1" pour confirmer chaque élément d'information introduit ait constitué une fréquente raison de rappel. Avant de rappeler, bon nombre des répondants qui ont corrigé leurs données avaient introduit le système IDTC pour la première fois ont rappelé deux fois plus souvent que les utilisateurs "expérimentés".

Tableau 3

Données de vérification des enregistrements – Erreur due au mode de collecte par IDTC avant et après les corrections décollant du contrôle

Heures de travail des employés à la production	Nombre d'employés rémunérés	Nombre d'employés de sexe féminin à la production	Nombre total d'employés	Erreur due au mode de collecte en pourcentage, avant les corrections décollant du contrôle	(Estimations de l'enquête)	Erreur due au mode de collecte en pourcentage, après les corrections décollant du contrôle	(Estimations de l'enquête)
				0.0	(.4)	0.0	(0.0)
				0.5	(.3)	0.0	(0.0)
				7.3	(5.2)	0.0	(0.0)
				7.3	(3.8)	0.0	(0.0)
				4.4	(3.7)	0.0	(0.0)

(N = 1,886 à 1,930 pour chaque question).

Le tableau 3 montre les effets possibles des divergences sur les éléments d'information de la CESS, effets déterminés en faisant la somme des écarts entre les valeurs obtenues par la méthode IDTC et celles figurant sur la formule, puis en divisant le résultat par la somme de ces dernières. Dans le cas de la CESS, les estimations publiées sont établies à l'aide d'un estimateur fondé sur une chaîne de rapports. Les estimations contenues dans le tableau 3 ne tiennent pas compte de cet estimateur. Elle permettent toutefois de mesurer indirectement l'erreur due au mode de collecte par IDTC en attachant les estimations de l'enquête. Au seuil de 5%, aucune des erreurs ne s'écarte de façon significative de zéro. Toutefois, les risques d'erreurs dues au mode de collecte semblent être plus élevés dans le cas du nombre d'employés à la production, de la rémunération et des heures de travail des employés à la production. Dans cette étude, on a surestimé le nombre des employés à la production et la rémunération de 7.3%, et les heures de travail de 4.4%.

Par ailleurs, presque toutes les réponses divergentes auraient été rejetées au contrôle (selon les paramètres utilisés dans la CESS) et corrigées. Comme l'illustre le tableau 3, les corrections effectuées au contrôle ont pour résultat d'annuler l'effet des réponses divergentes sur les données. On peut citer comme exemples de divergences importantes le cas d'un répondant qui a inscrit par erreur la rémunération à la place du nombre d'employés à la production, augmentant ainsi ce dernier de plus de 10,000, et celui de deux répondants qui ont inscrit le nombre d'heures de travail des employés à la production à la place du nombre d'employés à la production, augmentant ce dernier de plusieurs milliers. On relève de grossières divergences du même ordre dans les réponses aux questions portant sur la rémunération et sur les heures de travail, notamment le cas de répondants qui indiquent des cents au lieu d'arrondir au plus proche dollar.

### 5.2 Données enregistrées mécanographiquement

Le système IDTC permet d'évaluer les problèmes qu'éprouvent les répondants à utiliser ce mode de collecte. Il peut enregistrer le nombre de fois que les répondants réintroduisent leurs données, ainsi que le nombre de fois qu'une question est relue une deuxième ou une troisième fois avant d'obtenir une réponse. Il peut aussi conserver des renseignements sur les répondants qui rattrouche avant d'introduire des données. Les données enregistrées mécanographiquement ont été recueillies pour un échantillon total de 1,203 observations durant une période de trois mois, en 1990. Les données ont été recueillies auprès d'environ 474 répondants uniques, dont un bon nombre ont fourni des données pendant



Le deuxième type de divergence était l'inscription de chiffres additionnels ou, dans quelques cas, de trop peu de chiffres (18 erreurs sur 177 tombaient dans cette catégorie). Ce problème s'est surtout produit à la question sur la rémunération, où quatre répondants ont essayé d'indiquer les cents au lieu d'arrondir au dollar près. Plusieurs des mêmes répondants ont semblé inscrire une demi-heure, 50, comme nombre d'heures de travail des employés à la production, plutôt que d'arrondir. Dans le cas du troisième type de divergence, les chiffres introduits au moyen du téléphone à clavier étaient identiques à ceux figurant sur la formule, à une exception près. Le nombre inscrit incorrectement nous a laissé croire que le répondant avait peut-être éprouvé des problèmes avec le clavier (ses doigts ayant pu glisser sur la touche se trouvant directement à côté ou au-dessous de la touche correspondant au bon chiffre). Nous avons relevé 17 divergences de ce type. Pour ce qui est du quatrième type de divergence, on l'a détecté principalement à la question portant sur le nombre total d'employés: huit entreprises ont inscrit un "1" en réponse à cette question en utilisant la méthode IDTC, alors qu'un nombre plus élevé d'employés figurait sur leur formule d'enquête. Nous supposons que les répondants ont appuyé deux fois sur le "1" lorsqu'ils ont confirmé leur réponse à la question précédente (le mois).

Enfin, quelques répondants avaient corrigé certaines données sur leur formule, sans les corriger au moment de la déclaration au téléphone à clavier. Nous avons aussi relevé d'autres divergences que nous n'avons pu expliquer. En outre, plusieurs répondants ont transposé les chiffres déclarés ou les ont déclarés avec un décalage d'une catégorie ("autres raisons"). Dans la plupart des cas, il était difficile d'établir explicitement si les erreurs avaient été commises au moment de l'introduction des données ou au moment de la vérification, par suite d'une mauvaise compréhension de la question ou des chiffres lors de leur relecture par l'ordinateur. Bien que nous ayons soupçonné qu'elles avaient été commises au moment de l'introduction des données, il nous a été impossible de vraiment écarter la possibilité que les problèmes de compréhension aient eu une incidence, sauf dans le cas du deuxième type de divergence. Comme l'indique le tableau 2, les taux d'erreurs enregistrés pour les différentes questions de l'enquête variaient de 1,2% à 2,5%. De même, on observe un taux d'erreurs moins élevé pour les questions portant sur le nombre total d'employés, sur le nombre d'employés de sexe féminin et sur le nombre d'employés à la production que pour les questions portant sur la rémunération et sur les heures de travail. Cette situation n'a rien de surprenant, puisqu'il faut habituellement inscrire de quatre à six chiffres en réponse aux questions portant sur la rémunération et sur les heures de travail, comparativement à deux ou trois dans le cas des autres questions. Ainsi, les séries de chiffres plus longues posent davantage de problèmes aux répondants. Ces problèmes peuvent être attribuables au fait que les répondants éprouvent des difficultés à introduire les données au clavier ou à se rappeler de séries de chiffres plus longues durant la validation, ou encore au fait qu'ils ne sont pas motivés à apporter les corrections nécessaires.

Tableau 2

Données de vérification des enregistrements - Nombre de divergences et pourcentage d'erreurs par élément d'information

Heures de travail des employés à la production	Nombre d'employés à la production	Nombre d'employés de sexe féminin	Nombre d'employés à la production	Rémunération	Total
23	29	28	48	49	177
Erreur en pourcentage	1.2	1.5	1.5	2.5	1.8
(Estimations d'enquête)	(.2)	(.3)	(.3)	(.4)	(.3)

(N = 1,930 pour chaque question).

5. RÉSULTATS

5.1 Données de vérification des enregistrements

Lorsque nous avons choisi de demander aux entreprises qui répondaient par IDTC de renvoyer leurs formules d'enquête, nous nous sommes d'abord demandé combien de répondants avaient vraiment utilisé les formules. Nous supposons qu'une des sources de l'erreur due au mode de collecte était le fait que les répondants ne remplissaient pas la formule avant d'introduire leurs données au téléphone à clavier, ce qui plaçait un fardeau accru sur leur mémoire. Selon nous, ces répondants couraient plus de risques d'introduire et/ou de confirmer des données inexactes. C'est pourquoi, lorsque nous avons demandé aux répondants de nous renvoyer les formules d'enquête, nous leur avons indiqué qu'ils devaient nous les retourner peu importe s'ils les avaient remplies ou non. Toutefois, parmi les 96% d'entreprises qui ont renvoyé leur formule d'enquête, une seule a retourné une formule vierge. Bien qu'il soit possible que les non-répondants travaillaient de mémoire, la plupart des répondants avaient rempli leur formule, ce qui nous amène à croire que les problèmes de mémoire dus au fait qu'on n'avait pas rempli la formule ne constituaient pas une source majeure d'erreurs.

En comparant les données obtenues par IDTC et celles figurant sur les formules d'enquête, nous avons détecté et codé certaines divergences. Les données figurant sur la formule d'enquête sont celles que nous aurions obtenues si les répondants avaient participé à l'enquête par la poste. On trouve les résultats de cette comparaison au tableau 1. Le premier type de divergence était le suivant: les données IDTC indiquaient qu'on n'avait pas répondu à une question, mais une réponse figurait sur la formule d'enquête de l'entreprise. Cette non-réponse à certaines questions intervenait pour le plus grand nombre de divergences (82 sur 177) et était répartie assez uniformément sur toutes les questions de l'enquête.

Par ailleurs, la non-réponse aux questions variait selon le mois et l'entreprise. Ainsi, le taux de non-réponse était de 40% supérieur le premier mois où une entreprise déclarait les données à l'aide de la méthode IDTC. De plus, certaines entreprises éprouvaient plus de difficultés que d'autres (comme l'indique le fait qu'elles aient laissé deux questions ou plus sans réponse). Presque la moitié des cas de non-réponse à certaines questions ont été enregistrés auprès de 18 entreprises, au moment où elles ont commencé à utiliser la méthode IDTC ou peu après. Cette constatation nous a permis de conclure que la première utilisation de la méthode soulevait certains problèmes susceptibles de s'atténuer avec le temps. Comme le problème était circonscrit à un petit groupe d'entreprises, nous croyons qu'il découlait d'un manque de connaissance des processus automatisés. Les autres cas de non-réponse ne suivaient aucune tendance identifiable; nous avons présumé que certains répondants oubliaient tout simplement de répondre à une question, peut-être en raison d'une source de distractions dans leur bureau, et passaient à la question suivante.

Tableau 1

Données de vérification des enregistrements – Nombre et type de divergences obtenues avec la méthode IDTC

Non-réponse à une question par IDTC	82
1 à 2 chiffres manquants ou de trop	18
Glissement sur le clavier	17
Divergences/confirmation "1", erreur "0"	14
Formule corrigée, mais pas les données IDTC	12
Aucune raison apparente pour les erreurs	26
Autres raisons	8
Total	177



il nous fallait étudier les erreurs et les problèmes découlant de la complexité de la tâche et de sa compréhension par les répondants, y compris la possibilité d'améliorer le rendement des répondants avec le temps.

Nous avons décidé d'évaluer les problèmes soulevés par l'IDTC et l'erreur due à ce mode de collecte à partir de trois sources de données différentes, portant toutes sur un groupe commun d'environ 465 entreprises de la Pennsylvanie. Ces entreprises déclaraient leurs données d'enquête mensuelles par IDTC au laboratoire des techniques de collecte automatisée (ACT), au siège national du BLS, à Washington. Un petit nombre de ces entreprises ont commencé à déclarer leurs données au laboratoire ACT par IDTC en avril 1989 et les autres se sont graduellement jointes à ce groupe de départ d'un mois à l'autre, jusqu'en novembre 1989. La plupart des entreprises ont continué de déclarer des données au laboratoire ACT jusqu'en avril 1990. La majorité de ces entreprises ont passé de la méthode de collecte des données par la poste à la méthode IDTC.

Le premier ensemble de données se divise en deux composantes: les données IDTC qui ont été enregistrées mécanographiquement d'avril à décembre 1989 et les mêmes données consignées par les entreprises sur une formule d'enquête. Tous les répondants reçoivent chaque année une formule d'enquête sur laquelle on leur demande de consigner leurs données pour chaque mois. Les entreprises qui répondent par la poste remplissent la formule chaque mois et l'envoient par la poste à l'organisme d'État responsable de la sécurité de l'emploi. Ce dernier enregistre les données, puis renvoie la formule par la poste pour la collecte des données du mois suivant. Les entreprises qui répondent par ITAO et par IDTC reçoivent la formule par la poste, mais elles ne la renvoient pas. Toutefois, nous avons demandé aux répondants employant la méthode IDTC de renvoyer leur formule d'enquête de 1989 et nous avons enregistré un taux de retour de 96%. Nous avons ensuite comparé les données obtenues par IDTC et celles qui avaient été consignées sur la formule afin de détecter les divergences entre les deux. Les deux sous-ensembles de données comprennent 1,930 observations faites sur une période de neuf mois. Étant donné qu'on a introduit lentement et progressivement la méthode IDTC auprès des entreprises, le nombre d'observations par entreprise varie. Nous disposons de données couvrant une période de six à neuf mois pour environ 75 entreprises, une période de quatre à six mois pour environ 200 entreprises et une période de deux à trois mois pour environ 190 entreprises. Ces données sont désignées sous le nom de données de vérification des enregistrements.

Le deuxième ensemble de données comprend des renseignements enregistrés mécanographiquement qui se rapportent au rendement des répondants durant l'appel téléphonique IDTC. En janvier 1990, on a reprogrammé les ordinateurs utilisés pour l'IDTC afin qu'ils totalisent et enregistrent automatiquement le nombre de fois qu'une question a été répétée en raison d'une non-réponse (catégorie des questions répétées), le nombre de fois qu'un répondant a réintroduit des données (catégorie des données réintroduites) pour chaque question, ainsi que le nombre de fois qu'une entreprise a appelé et raccroché avant d'introduire des données. Malheureusement, les seules questions pour lesquelles il a été possible d'effectuer des relevés distincts pour les catégories des questions répétées et des données réintroduites sont celles concernant le mois et le nombre total d'employés. Dans le cas des questions portant sur le nombre d'employés de sexe féminin, sur le nombre d'employés à la production, sur la rémunération et sur les heures de travail, la structure du programme machine initial nous a forcés à combiner les données sur les répétitions de questions et sur les réintroductions de données. Ces données sont désignées sous le nom de données enregistrées mécanographiquement.

La troisième source de données est une enquête d'évaluation téléphonique qui a été réalisée de janvier à avril 1990 auprès des entreprises de la Pennsylvanie et qui portait sur leur expérience avec le système IDTC. Environ 411 entreprises ont répondu à l'interview, pour un taux de réponse de 88%. Les questions portaient sur des sujets tels que la qualité de la voix, le débit, les problèmes relatifs à la tâche, l'utilisation des fonctions du système, l'à-propos des instructions et une évaluation du système.



et le service téléphonique automatique. Bien que ces services permettent de gagner du temps et d'économiser de l'argent, ils peuvent parfois rebuter les utilisateurs. Par ailleurs, le système, la tâche ou le répondant peuvent tous être la source de problèmes et d'erreurs. Les problèmes liés au système engendrent principalement une erreur due à la non-réponse, tandis que la complexité de la tâche et la qualité de son exécution par le répondant sont la cause d'une erreur de mesure.

Quoi qu'il en soit, ils ne traitent pas directement des enquêtes, les ouvrages sur l'ergonomie indiquent plusieurs facteurs interdépendants pouvant contribuer aux erreurs d'exécution découlant de l'utilisateur d'une interface homme-machine. D'une part, il se peut que les répondants ne soient ni familiers ni à l'aise avec la technologie. D'après Waterworth (1984), le langage employé dans l'interface homme-machine est différent de la communication humaine étant donné que les actes sont exécutés suivant un ordre reflétant la logique de la programmation informatique. Comme il n'est pas facile d'adopter un mode de pensée épousant cette logique, les personnes ayant peu d'expérience de l'informatique peuvent éprouver de la difficulté à comprendre la tâche qu'elles doivent exécuter et à utiliser le système. D'autre part, la parole synthétique est plus difficile à comprendre que la parole naturelle et pose des exigences plus élevées sur la fonction traitement de la mémoire immédiate (Schwab, Nusbaum et Pisoni 1985). Ainsi, les problèmes de compréhension et de mémoire associés à ce mode de collecte des données peuvent être la source d'erreurs.

La parole synthétique comprend à la fois la parole numérisée, où une voix humaine est échantillonnée, codée numériquement et enregistrée, et la parole synthétisée suivant des règles, laquelle est produite à l'aide d'un texte servant d'intrant (Marics et Williges 1988). Le système IDTC utilise la parole numérisée, laquelle est moins difficile à comprendre que la parole synthétisée suivant des règles. Toutefois, la numérisation soulève des problèmes de compréhension du fait qu'elle provoque une distorsion de la parole naturelle (Cox et Coope 1981). Certains travaux de recherche ont toutefois démontré qu'il est possible d'améliorer la compréhension de la parole synthétique par la formation. Lors d'une expérience sur la perception de la parole de la parole synthétique par la formation. Schwab et ses collaborateurs (1985) ont découvert que le fait de recevoir une formation sur la parole synthétique permet d'améliorer la perception. Ainsi, la compréhension de cette parole peut s'améliorer à mesure qu'on acquiert de l'expérience dans l'utilisation du système. Le débit constitue un autre facteur qui peut influencer sur la compréhension. Marics et Williges (1988) ont découvert, en mesurant les erreurs de transcription et le temps de latence (temps écoulé entre la fin de la question et le début de la réponse correspondante), que le débit de la parole synthétique a une incidence considérable sur son intelligibilité. Toutefois, les sujets qui ont reçu des renseignements contextuels avant d'écouter la parole ont commis moins d'erreurs de transcription.

Par conséquent, les erreurs susceptibles de découler de l'utilisation d'une interface homme-machine peuvent être attribuables à un manque d'expérience avec la technologie et la tâche à exécuter, ainsi qu'à des problèmes de compréhension et de mémoire reliés au débit et à la clarté de la voix. Ces problèmes peuvent néanmoins être surmontés, puisqu'il a été démontré que, avec de l'expérience et une formation adéquate, il est possible d'améliorer le rendement.

#### 4. LES DONNÉES

Plusieurs objectifs ont été pris en considération au moment de l'évaluation des problèmes soulevés par l'IDTC et de l'erreur due au mode de collecte, ainsi que de la détermination des données à utiliser ou à recueillir. Premièrement, il était nécessaire de déterminer si des problèmes se posaient et, le cas échéant, à quelle étape du processus ils se posaient. Deuxièmement, nous estimions que les répondants devaient cerner et interpréter eux-mêmes les problèmes, mais nous voulions aussi obtenir des mesures indépendantes des évaluations des répondants. Troisièmement,

Grâce au système IDTC du BLS, il est possible:

- de repérer les entreprises répondantes légitimes d'après un appariement avec un fichier de numéros d'entreprise;

- de faire varier la série de questions selon l'activité économique exercée par l'entreprise; de repasser toutes les réponses afin que le répondant puisse les confirmer, à l'aide d'une voix simulée sur ordinateur (numérisée) (on demande aux répondants d'inscrire un "1" pour confirmer leur réponse ou un "0" pour la réintroduire);

- d'attendre deux secondes avant que les répondants inscrivent leur réponse, et d'attendre deux secondes entre les chiffres avant de considérer l'introduction des données comme étant terminée (on suppose aussi que l'introduction des données est terminée si toutes les positions de la zone ont été remplies; par exemple, la zone correspondant au "mois" comprend deux chiffres);

- de répéter chaque question jusqu'à trois fois (pour le numéro d'identification, le mois et le nombre total d'employés) ou de demander au répondant de confirmer qu'il n'a pas de réponse pour la question (dans le cas de tous les autres éléments d'information), lorsqu'un répondant ne confirme pas sa réponse ou ne donne pas de réponse dans les deux secondes suivant la lecture de la question;

- d'enregistrer la date, l'heure du début et de la fin de l'appel, et tous les éléments d'information (Werking et coll., 1988). Les répondants reçoivent par la poste des instructions sur l'IDTC dans lesquelles on leur indique la façon d'utiliser le système et on leur donne des exemples de dialogue entre l'ordinateur et le répondant, comme celui-ci:

Ordinateur:	
Inscrivez le nombre total	
Dans le cas de 25 employés, appuyez	
sur le 2 et le 5	
Vous avez inscrit 2, 5	
Appuyez sur 1 pour confirmer ou sur	
le 0 pour réintroduire des données.	

Dans les instructions, on indique aussi aux répondants qu'ils ont le choix d'utiliser le dîse ("##") pour signaler qu'ils ont terminé l'introduction des données pour une question, ce qui permet de réduire la durée de l'interview. En outre, avant de déclarer des données, les répondants peuvent téléphoner pour faire l'essai du système en utilisant un numéro d'identification spécial qui leur est délivré à des fins d'essai.

Un intervieweur communique par téléphone avec les répondants employant la méthode IDTC durant leur premier mois d'utilisation du système pour savoir s'ils ont éprouvé des problèmes et, au besoin, pour leur prodiguer des conseils. Chaque mois, les répondants reçoivent une carte de rappel par la poste et, s'ils n'ont pas répondu d'eux-mêmes avant une date préalable, un appel de sollicitation leur demandant d'entrer en communication avec le système IDTC le plus tôt possible. En général, les données ne sont pas recueillies au moment de l'appel de sollicitation.

### 3. ERREUR DE MESURE DÉCOULANT DE L'UTILISATION D'UNE INTERFACE HOMME-MACHINE

Il existe peu de précédents en matière d'utilisation d'un système IDTC pour répondre aux questions d'une enquête. Toutefois, la reconnaissance des données introduites au moyen d'un clavier est très répandue dans le cas de services tels que les virements bancaires électroniques



les délais de réponse, le coût de l'utilisation exclusive de la méthode ITAO pour la CESS ne peut pas être absorbé à l'intérieur du budget actuel affecté à cette enquête. Des travaux de recherche sur l'IDTC ont été réalisés depuis 1986 afin d'élaborer une autre méthode de collecte des données qui offrirait les mêmes améliorations au chapitre du rendement que la méthode ITAO, mais à un coût moindre (Ponikowski, Copeland et Meilly 1989). Dans leur article, Werking et Clayton font une analyse plus poussée de l'utilisation des méthodes ITAO et IDTC pour la CESS.

Le BLS a effectué récemment des essais sur le système IDTC afin d'obtenir des données sur les délais de réponse, sur le coût de la collecte et sur les taux de rejet au contrôle. Ces essais ont montré que la collecte par IDTC utilisant l'ITAO comme méthode d'appoint est aussi efficace que la méthode ITAO et qu'elle permet de respecter des délais de réponse aussi serrés (Werking, Tupek et Clayton 1988). De plus, l'IDTC permet de réduire considérablement les coûts afférents à l'enquête. Selon les estimations actuelles, les coûts unitaires mensuels de la collecte par IDTC sont d'environ 30% inférieurs à ceux de la collecte par la poste, tandis que la collecte par ITAO est 20% plus coûteuse que la collecte par la poste (Clayton et Harrell 1990). On étend maintenant l'utilisation de la méthode de collecte par IDTC à une plus grande partie de la population de la CESS, en commençant par les entreprises comptant le plus grand nombre d'employés. Les travaux de recherche visant à trouver des moyens d'améliorer le système IDTC se poursuivent, même si les répondants ont réservé un accueil très favorable à celui-ci.

Les travaux de recherche auxquels le BLS procède actuellement ont pour objet de cerner les problèmes que les répondants éprouvent à utiliser le système IDTC et de mesurer les erreurs dues à ce mode de collecte des données. Les résultats de ces travaux indiquent les points qu'il faut améliorer pour faciliter la tâche des répondants ainsi que pour réduire les erreurs. La section 2 contient des renseignements généraux sur la méthode IDTC et sur le fonctionnement du système IDTC du BLS. Dans la troisième section, nous traitons des erreurs qui peuvent être reliées à l'utilisation de l'IDTC, qui nécessite une interface homme-machine, comme mode de collecte des données. Dans la section 4, nous décrivons les trois sources de données utilisées aux fins de l'étude. Les problèmes et les erreurs sont analysés à l'aide des résultats d'une enquête de vérification des enregistrements, de données enregistrées mécanographiquement, ainsi que des résultats d'une enquête d'évaluation auprès des répondants. La section 5 fait état des méthodes utilisées, des analyses effectuées et des résultats obtenus à partir de chacune des trois sources de données. Enfin, dans la section 6, nous présentons une évaluation globale des erreurs de mesure dues à ce mode de collecte, certaines suggestions en vue d'améliorer le système ainsi que certaines observations quant à l'utilisation éventuelle de ce système pour d'autres enquêtes.

## 2. INTRODUCTION DES DONNÉES AU MOYEN D'UN TÉLÉPHONE À CLAVIER

La principale raison pour laquelle on envisage d'utiliser le mode de collecte par IDTC est afin de réduire le coût de collecte des données par ITAO, tout en maintenant l'actualité et la qualité des données au niveau actuel. La CESS semblait bien se prêter à la collecte des données par IDTC, puisque seulement cinq ou six éléments d'information numériques y sont recueillis chaque mois. Ces éléments d'information comprennent: le nombre total d'employés, le nombre d'employés de sexe féminin, le nombre d'employés à la production, la rémunération des employés à la production, le nombre d'heures de travail des employés à la production et, dans le cas de certaines activités économiques, le nombre d'heures supplémentaires ou les commissions versées aux employés. Les entreprises sont priées de déclarer les totaux relatifs à chaque élément d'information pour la période de paye dans laquelle tombe le 12<sup>e</sup> jour du mois. On demande aussi aux répondants employant la méthode IDTC d'indiquer le numéro d'identification de leur entreprise et le mois pour lequel ils fournissent des données.



## Fiabilité des données introduites au moyen d'un téléphone à clavier

POLLY A. PHIPPS et ALAN R. TUPEK<sup>1</sup>

### RÉSUMÉ

Le Bureau of Labor Statistics a mis en oeuvre, dans le cadre d'une enquête mensuelle auprès des entreprises, une méthode de collecte électronique des données utilisant la reconnaissance des données introduites au moyen d'un téléphone à clavier. Le système d'introduction des données au moyen d'un téléphone à clavier (IDTC) utilise des expressions numérisées pour demander aux participants à l'enquête de répondre aux questions en appuyant sur les boutons d'un poste téléphonique à clavier numérique. Le système IDTC permet de réduire de façon substantielle les coûts d'enquête par suite de l'élimination de nombreuses activités à prédominance de main-d'oeuvre. On en sait toutefois très peu au sujet des erreurs de mesure découlant de l'utilisation de ce mode de collecte. Dans la présente étude, nous évaluons les erreurs inhérentes à l'emploi de la méthode IDTC à partir de trois sources de données; il est ainsi possible d'analyser les erreurs reliées à des aspects précis de l'interface homme-machine. De plus, nous faisons état de certaines caractéristiques de conception du système qui ont des répercussions sur les erreurs dues à ce mode de collecte. Nous terminons en traitant de l'incidence de nos constatations sur d'autres enquêtes.

**MOTS CLÉS:** Mode de collecte des données; interface homme-machine; auto-interview assistée par ordinateur.

### 1. INTRODUCTION

Le Bureau of Labor Statistics (BLS) des États-Unis diffuse chaque mois des estimations de l'emploi pour l'ensemble des États-Unis. Ces estimations sont établies à partir des données recueillies au moyen d'une enquête réalisée auprès de 350,000 entreprises, la Current Employment Statistics Survey (CESS), qui fournit l'une des premières mesures mensuelles de l'état de santé de l'économie américaine. Toutefois, les estimations provisoires de l'enquête sont établies à partir de données ne provenant que d'environ la moitié des entreprises participant à l'enquête, des estimations révisées étant produites deux mois après la diffusion initiale des données aux médias. Il peut arriver, en raison du faible taux de réponse enregistré au moment de l'établissement des estimations provisoires, que ces dernières doivent faire l'objet de révisions importantes. En 1983, le BLS a entrepris des travaux de recherche sur l'utilisation de techniques de collecte automatisée des données afin d'assurer un meilleur respect des délais de réponse et de réduire l'importance éventuelle des révisions nécessaires.

Traditionnellement, les questionnaires de la CESS étaient recueillis par la poste, par l'intermédiaire des organismes d'État responsables de la sécurité de l'emploi. Des essais réalisés entre 1984 et 1986, au cours desquels on a remplacé la méthode de collecte par la poste par une interview téléphonique assistée par ordinateur (ITAO), ont démontré que la méthode ITAO constituait un moyen efficace de réduire les délais de réponse (Werking, Tupak, Ponikowski et Rosen 1986). Les taux de réponse obtenus pour les estimations provisoires en utilisant la méthode ITAO se sont situés entre 85% et 90%, comparativement à 45% à 50% avec la méthode de collecte des données par la poste. Bien que la méthode ITAO ait permis de réduire

<sup>1</sup> Polly A. Phipps, Office of Employment and Unemployment Statistics, Bureau of Labour Statistics, Room 2821, 441 G Street N.W., Washington, D.C. 20212; Alan R. Tupak, Office of Employment and Unemployment Statistics, Bureau of Labor Statistics, Room 2919, 441 G Street N.W., Washington, D.C. 20212.

## BIBLIOGRAPHIE

- CLAYTON, R.L., et HARRELL, L., Jr. (1989). Developing a cost model of alternative data collection methods: MAIL, CATI et TDE. *Proceedings of the Section on Survey Research Methods, American Statistical Association*.
- CLAYTON, R.L., et WINTER, D.L.S. (1990). Speech data entry: Results of the first test of voice recognition for data collection. *Proceedings of the Section on Survey Research Methods, American Statistical Association*.
- JURAN, J.M., GRYNALIS, F.N., Jr., et BINGHAM, R.S., Jr., eds. (1979). *Quality Control Handbook*, troisième édition. McGraw Hill.
- GROVES, R.M.J., et coll., eds. (1988). *Telephone Survey Methodology*. New York: John Wiley and Sons.
- OFFICE OF MANAGEMENT AND BUDGET (1988). *Quality in Establishment Surveys*. Statistical Policy Working Paper 15.
- OFFICE OF MANAGEMENT AND BUDGET (1990). *Computer Assisted Survey Information Collection*. Statistical Policy Working Paper 19.
- PONIKOWSKI, C., et MEELY, S. (1988). Use of touchtone recognition technology in establishment survey data collection. Presented at the First Annual Field Technologies Conference, St. Petersburg, Florida.
- WERKING, G.S., TUPEK, A.R., PONIKOWSKI, C., et ROSEN, R. (1986). A CATI feasibility study for a monthly establishment survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 639-643.
- WERKING, G.S., et TUPEK, A.R. (1987). Modernizing the Current Employment Statistics Program. *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, 122-130.
- WERKING, G., TUPEK, A., et CLAYTON, R. (1988). CATI and touchtone self-response applications for establishment surveys. *Journal of Official Statistics*, 349-362.

études pendant plusieurs mois. En appliquant ces taux projetés de réduction de l'ampleur des révisions aux révisions effectuées au cours des 5 dernières années, on a découvert que 97% des révisions se situaient alors au-dessous du seuil de 50,000 comparativement à seulement 60% actuellement. Cette réduction considérable de la taille de l'échantillon visé par le passage à la méthode de l'ITAO/de l'IDTC a rendu possible l'adoption d'un calendrier de mise en oeuvre accélérée permettant de maintenir les coûts de conversion au niveau minimal nécessaire pour éliminer les révisions d'importance des estimations provisoires. Le Bureau entreprendra la mise en oeuvre des méthodes de collecte par ITAO/par IDTC dans 25 Etats en 1991 et prévoit étendre l'utilisation de ces méthodes à tous les Etats en 1992. Grâce à la mise en oeuvre de ces nouvelles méthodes de collecte, le Bureau sera en mesure de contribuer à la solution d'un des problèmes de qualité les plus épineux et les plus importants pour les utilisateurs des données de la CESS.

#### 4. RÉSUMÉ

Les années quatre-vingt ont été une période de changements multiples pour les organismes statistiques. Certains de ces changements ont été le résultat de nos réalisations, tandis que d'autres, plus subtils, sont venus modifier le cadre général dans lequel nous réalisons nos enquêtes.

Au cours des années quatre-vingt, les utilisateurs sont devenus beaucoup plus conscients de la qualité des données et prompts à cerner et à signaler les limites de nos produits. Sans que la qualité de nos produits ne se soit nécessairement détériorée, les attentes des utilisateurs sur le plan de la qualité et de l'aptitude à l'usage se sont considérablement accrues. À titre d'exemples statistiques, nous nous devons de relever ce défi afin d'être en mesure de maintenir notre crédibilité auprès des utilisateurs. Nous avons également assisté au cours de la dernière décennie à de spectaculaires percées technologiques et, en particulier, dans le domaine de la micro-informatique. La nouvelle technologie a offert aux organismes statistiques de nombreuses occasions nouvelles d'améliorer la qualité et le contrôle de la collecte des données, par exemple: l'ITAO, l'AIMAO, l'IDTC, la RP et le télécopieur. Certaines de ces techniques permettent d'améliorer la qualité et le contrôle tout en réduisant les coûts de fonctionnement courants. Il est d'ailleurs fort possible que la prochaine décennie nous offre des occasions encore meilleures d'utiliser les nouvelles technologies pour améliorer la qualité et réduire les délais de collecte des données tout en diminuant les dépenses.

Lorsque nous examinons l'état de nos programmes statistiques, nous découvrons fréquemment un cadre de mise en oeuvre très rigide. Souvent, la méthode de collecte des données utilisée pour nos enquêtes n'a jamais été modifiée depuis la mise en oeuvre initiale de ces dernières. Les hypothèses que nous utilisons au sujet des coûts de collecte des données et nos analyses des coûts sont d'ordinaire nettement périmees et s'appuient sur des approches souvent simplistes. Comme la collecte des données intervient généralement pour la majeure partie des coûts d'une enquête, elle est d'ordinaire bien ancrée dans la structure organisationnelle de l'organisme et il peut être très difficile de la restructurer afin de permettre d'effectuer des modifications d'envergure. C'est à l'intérieur de ce cadre de réalisation des enquêtes que nous devons relever les défis et saisir les occasions que nous offriront les années quatre-vingt-dix.

Au cours de la prochaine décennie, les organismes statistiques devront relever le triple défi: d'être sensibles à l'évolution des besoins des utilisateurs en matière de qualité des données; de tout faire pour que leurs travaux de recherche suivent l'évolution rapide de la technologie et des méthodes de collecte automatisée des données; enfin, peut-être par dessus tout, de continuer à trouver des façons d'incorporer les résultats des travaux de recherche couronnés de succès aux programmes courants.

Ces défis détermineront la future position concurrentielle de nos programmes et de nos organismes en termes de coûts et de qualité.



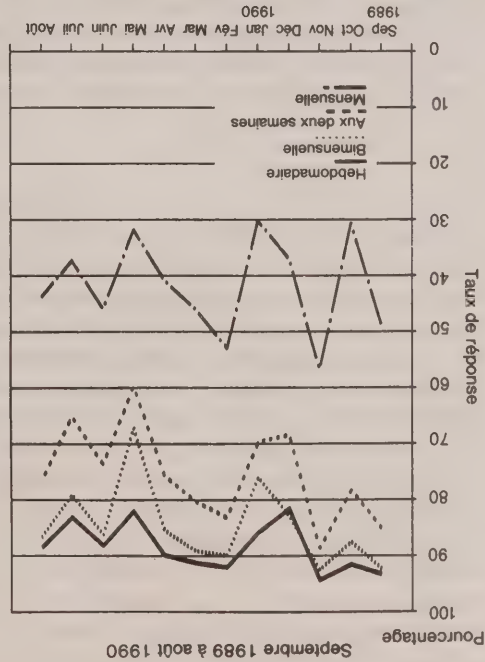


Figure 3. CESS, rendement de l'ITAO à la première date limite, selon la durée de la période de paie

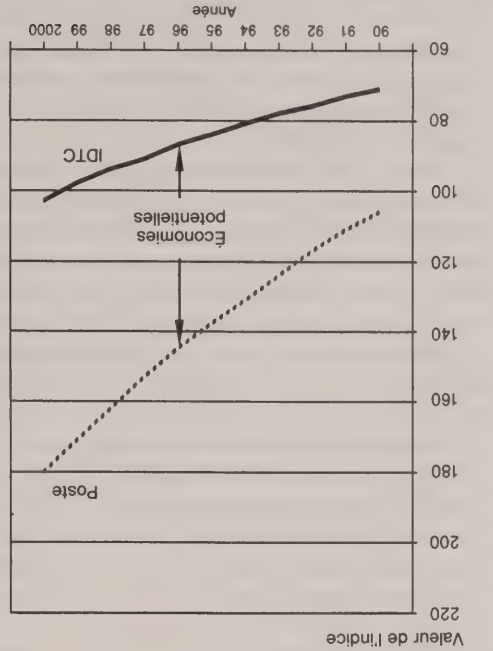


Figure 5. Coûts estimatifs par unité, selon le mode: de 1990 à 2000  
Nota: Projections non officielles du BLS

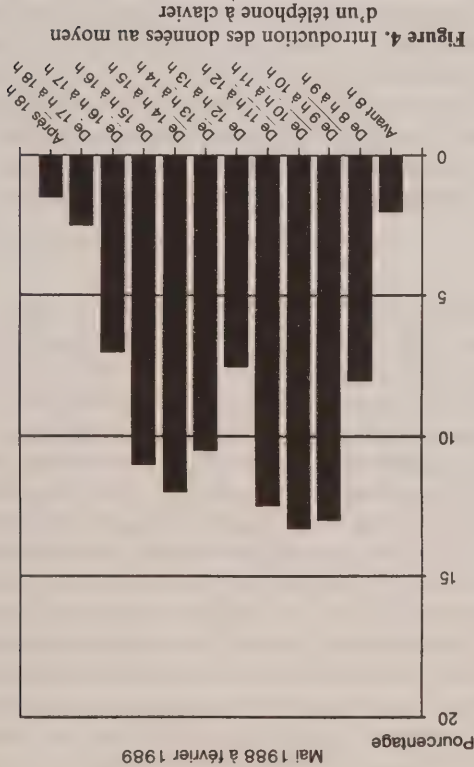
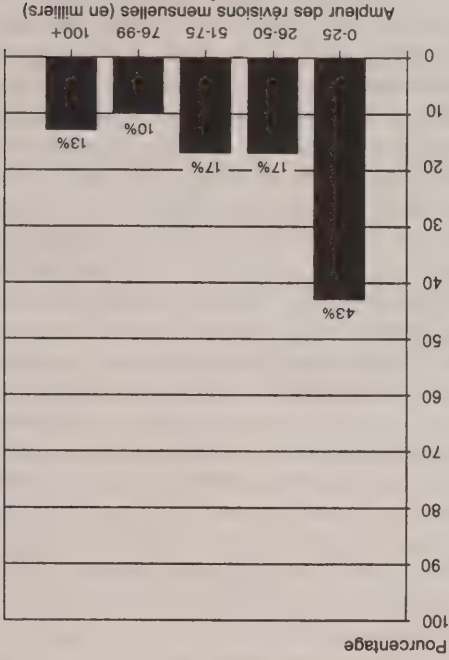


Figure 6. Répartition des révisions selon leur ampleur (De 1985 à 1989)



garder à l'esprit que le système de collecte étudié était en définitive celui des Etats et qu'il devait donc être conçu pour s'intégrer facilement à leurs mécanismes d'enquête et avoir l'incidence organisationnelle la plus faible possible. À cette fin, les systèmes d'ITAO et d'IDTC ont pu être modifiés pendant toute la durée des travaux. Chaque nouvelle version des systèmes tenait compte du plus grand nombre possible de suggestions et d'exigences des Etats. On peut dans une large mesure attribuer la réussite des travaux d'élaboration à l'ouverture d'esprit et à la persévérance dont ont fait preuve les 14 Etats participant aux travaux en ne cessant de formuler de nouvelles recommandations visant à améliorer les systèmes et les procédures. À la fin, la méthode de l'ITAO s'était transformée d'une simulation malhabile d'interviews de type enquête-ménage en une technique rapide et efficace à "écrans" et "fenêtres" bien adaptée à la saisie et au contrôle de données économiques longitudinales. Ainsi, à la conclusion des travaux de recherche, les systèmes et procédures avaient fait l'objet de nombreux essais et améliorations dans un large éventail d'Etats. Cette façon d'aborder les essais a permis de développer, au niveau des Etats, une confiance solide dans les méthodes et les systèmes proposés. Cette caractéristique s'est révélée essentielle pour le respect du calendrier de mise en oeuvre rapide de ces méthodes de collecte d'avant-garde proposé par le Bureau.

### 3.2 Approche et incidence

L'objectif principal de la proposition de mise en oeuvre était de maintenir au-dessous d'un seuil limite l'ampleur des révisions apportées aux estimations provisoires. Au cours des 5 dernières années, dans environ 40% des cas, la différence entre les estimations provisoires et les estimations définitives s'élevait à plus de 50,000 et dans 13% des cas à plus de 100,000 (c.-à-d. qu'il s'agissait de révisions importantes) (figure 6). L'étude de mise en oeuvre avait pour objectif de repérer un ensemble minimal d'entreprises donnant des réponses tardives qui, si on pouvait obtenir leurs réponses avant la date limite pour la diffusion, permettraient de maintenir l'ampleur des révisions au-dessous d'un seuil jugé acceptable (modification s'élevant à moins de 50,000 par rapport au chiffre des fluctuations mensuelles relatives à l'emploi). Bien sûr, il aurait été possible de demander à l'ensemble des 175,000 entreprises donnant des réponses tardives de produire leurs déclarations à l'aide des nouvelles méthodes de collecte, mais on a jugé que la mise en oeuvre d'une telle solution serait longue et coûteuse. Bien qu'il ait été nécessaire de maintenir l'ampleur des révisions au-dessous d'un seuil acceptable (c.-à-d. que l'objectif visé n'était pas d'éliminer toutes les révisions), il fallait également s'assurer d'atteindre cet objectif dans les meilleurs délais (c.-à-d. en changeant la méthode de collecte utilisée pour le plus petit nombre possible d'entreprises afin de permettre le maintien des révisions au-dessous du seuil de 50,000).

Contrairement aux enquêtes-ménages, les enquêtes auprès des établissements utilisent en général une pondération différentielle pour les diverses unités de collecte, le plan de sondage prévoyant que les unités très importantes doivent faire partie de l'échantillon. Dans le plan de sondage de la CESS, les unités comptant 100 employés ou plus ne forment que 20% de l'échantillon (c.-à-d. qu'il y en a 75,000), mais interviennent pour plus de 83% de l'emploi dans l'échantillon non pondéré. Comme ces unités ont tendance à afficher un taux de réponse beaucoup moins élevé pour les estimations provisoires, toute divergence entre la tendance de l'emploi chez les unités donnant une réponse tardive et la même tendance chez les unités répondant dans les délais fixés est susceptible de nécessiter une révision substantielle des estimations établies à partir de l'échantillon. Des études ont été réalisées afin d'évaluer l'incidence des unités comptant 100 employés ou plus sur les estimations provisoires. Pour vérifier l'incidence des gros employeurs, les unités comptant 100 employés ou plus donnant des réponses tardives ont été incluses dans l'échantillon initial utilisé pour établir les estimations provisoires et les estimations ont été recalculées. On a ensuite comparé ces nouvelles estimations aux estimations provisoires initiales afin de déterminer l'incidence des unités comptant 100 employés ou plus sur les révisions. Les résultats ont indiqué qu'entre une demi et deux tiers de la révision ou plus sur les révisions. Les résultats ont indiqué qu'entre une demi et deux tiers de la révision était attribuable à ces unités. On a ensuite obtenu des résultats similaires en effectuant les mêmes



causer très peu de perturbations dans la production des déclarations mensuelles. Selon les entretiens de suivi réalisés par le Bureau auprès des répondants pour lesquels la méthode de collecte des données est maintenant l'IDTC, ces derniers éprouvent très peu de difficulté à s'adapter à cette nouvelle méthode de déclaration. Virtuellement tous les répondants ont rempli leur première déclaration mensuelle par IDTC avec précision et sans aide et de nombreux répondants ont mentionné comment il était facile de produire une déclaration à l'aide de cette méthode. (3) On peut considérer que l'IDTC constitue une méthode de remplacement fiable pour la collecte des données d'enquête. Au cours des quatre dernières années de collecte, on a enregistré ni défaillance majeure du matériel ni interruption du processus de collecte. Au besoin, les défaillances mineures du matériel ont pu facilement être résolues à l'aide d'un micro-ordinateur de secours. En outre, le processus de collecte par IDTC dans les États prévoit une option de renvoi automatique des appels, destinée à réacheminer les appels à un site central en cas de défaillance majeure du système utilisé à l'échelle de l'État.

### 3. MISE EN ŒUVRE

#### 3.1 Principales difficultés

À la fin de 1989, le Bureau avait terminé un programme de recherche couronné de succès lui ayant permis d'assurer le maintien d'un rendement élevé tout au long des 7 années qu'avaient duré les travaux. Toutefois, il y a une différence importante entre la réalisation d'un programme de recherche réussi et la mise en œuvre de nouvelles méthodes à l'échelle réelle. Bien qu'on ait recueilli des données auprès de plus de 10,000 unités au cours de la mise à l'essai de ces nouvelles techniques, ces unités représentaient moins de 3% de l'échantillon de la CESS. L'adoption des modifications proposées pour la collecte mensuelle, après d'un échantillon de 350,000 unités, de données ayant été recueillies pendant plus d'un demi-siècle à l'aide d'un système décentralisé de collecte par la poste administré par les États nécessite non seulement l'expression d'un besoin manifeste de changement de la part des utilisateurs, mais aussi l'existence d'un soutien très large à l'échelle du pays, des régions et des États.

En l'occurrence, les besoins des utilisateurs avaient commencé à évoluer dès le début des années quatre-vingt. Au cours de cette décennie, l'économie des États-Unis a connu la plus longue période de croissance soutenue en temps de paix de son histoire, le nombre des emplois créés s'établissant à plus de 19 millions et les taux de chômage atteignant leur plus bas niveau depuis le début des années soixante-dix. Au milieu des années quatre-vingt, la politique économique était clairement axée sur l'établissement d'une croissance économique non inflationniste. Les observateurs examinaient d'un oeil attentif les données mensuelles de la CESS sur la croissance de l'emploi et sur les salaires afin d'y détecter tout signe de pression inflationniste par les coûts résultant de l'existence d'une forte croissance de l'emploi pendant une période de faible chômage. Cette utilisation et cette visibilité accrues des données mensuelles se sont accompagnées d'un accroissement correspondant de la frustration manifestée par les utilisateurs à l'égard des importantes révisions dont les estimations provisoires faisaient périodiquement l'objet. Bien que les estimations provisoires de la CESS aient toujours fait l'objet de révisions importantes et que l'importance de ces révisions ait diminué au fil des ans, l'apport de révisions supérieures à 100,000 aux estimations provisoires de l'évolution de l'emploi d'un mois à l'autre était maintenant considéré comme inacceptable. La demande des utilisateurs pour des estimations provisoires plus précises devait amener le Bureau à élaborer des propositions en vue de l'application des méthodes de collecte automatisée par ITAO et par IDTC dans le cadre de la plus importante enquête mensuelle réalisée par l'administration des États-Unis.

Bien que la demande des utilisateurs constitue un critère de première importance, des modifications de cet ordre ne sauraient être entreprises sans le soutien inconditionnel des États chargés de la collecte des données. Tout au long du programme de recherche, les chercheurs devaient



L'IDTC possède, sur le plan de la rentabilité et de la qualité, un avantage certain par rapport à la collecte par la poste du fait qu'elle facilite grandement le suivi téléphonique dans les cas de non-réponse. Grâce à l'IDTC, il est possible de produire une liste exacte et à jour des répondants n'ayant pas encore transmis leurs données par téléphone, puis d'utiliser cette liste pour effectuer de brefs appels afin de rappeler à ces personnes qu'elles doivent communiquer leurs données. Lorsque les données étaient recueillies par la poste, il était peu commode d'assurer le suivi téléphonique des cas apparents de non-réponse puisque les fonctionnaires de l'Etat ne savaient pas si la formule du répondant avait été remplie ou non, si elle avait été envoyée par la poste, ou encore si elle en était au stade de l'enregistrement ou à celui de la saisie; en outre, les répondants ayant récemment envoyé leur formule par la poste avaient tendance à s'offenser de recevoir un rappel au sujet d'une activité qu'ils considéraient terminée. Comme la participation à l'enquête était volontaire et qu'on ignorait si le répondant avait rempli son questionnaire ou non, ce n'est qu'aux unités critiques (gros employeurs) qu'on téléphonait pour leur rappeler de communiquer leurs données.

L'utilisation de la méthode de l'IDTC n'a permis de réaliser aucune économie sensible au chapitre de la conciliation, puisque le nombre de rejets au contrôle s'est établi à peu près au même niveau pour la collecte par IDTC que pour la collecte par la poste. Il en a été de même pour les cartes postales de rappel, le nombre de cartes envoyées aux établissements donnant une réponse tardive quand on utilisait la méthode de collecte par la poste étant à peu près égal au nombre de cartes de préavis envoyées aux répondants pour leur rappeler d'introduire leurs données au moyen d'un téléphone à clavier à la date prévue quand on employait la méthode de collecte par IDTC.

Sur le plan des catégories de dépenses non salariales, les frais d'affranchissement (actuellement 58 cents par unité) sont remplacés par la somme des frais d'appel et de la fraction non amortie du coût de l'appareil utilisé pour l'IDTC (somme s'élevant actuellement à 46 cents par unité). Les frais d'affranchissement augmentent constamment avec une hausse annuelle de l'ordre de 5% environ (figure 2). Cette majoration est causée par un accroissement annuel des frais de main-d'oeuvre (+ 5,7%) et par le prix de l'essence qui augmente aussi de façon générale, les frais de main-d'oeuvre intervenant pour plus de 80% des frais totaux d'affranchissement. Par contraste, pour la méthode de l'IDTC, on a assisté au cours des dernières années à une diminution des frais d'appel (- 1,7%) ainsi que du prix d'achat des micro-ordinateurs (- 19,5%).

Sauf pour ce qui est de l'obligation nouvelle de téléphoner aux répondants pour leur rappeler de communiquer leurs données, dans les cas de non-réponse complète, on peut démontrer que le passage de la collecte par la poste à la collecte par IDTC permet de réaliser des économies. Toutefois, il est peut-être encore plus important de noter que, selon une projection sur une période de 10 ans des coûts afférents à l'utilisation de ces deux méthodes de collecte (figure 5), ces économies augmenteront considérablement. On tentera d'utiliser les économies réalisées à l'avenir grâce à l'utilisation de l'IDTC pour compenser les frais relatifs à l'obligation de téléphoner aux répondants afin de leur rappeler de communiquer leurs données dans les cas de non-réponse complète.

Cet examen du rendement et de l'analyse des coûts nous permettent de tirer plusieurs conclusions importantes relatives à la déclaration des données à l'aide de la méthode de l'IDTC. (1) La collecte des données par la poste ne constitue plus la méthode de collecte la moins coûteuse dont disposent les organismes statistiques. Non seulement les percées technologiques importantes des années quatre-vingt sur le plan de la collecte téléphonique automatisée permettent-elles d'abaisser les coûts d'utilisation de cette technique au-dessous de ceux de la collecte par la poste, mais elles permettent aussi de réduire les délais de collecte et d'exercer un meilleur contrôle sur le processus de collecte. En outre, au cours des 5 à 10 prochaines années, l'accroissement des frais de main-d'oeuvre et d'affranchissement relatifs à la collecte des données par la poste rendra cette dernière encore moins concurrentielle par rapport aux nouvelles méthodes à haute technologie et à faible intensité de main-d'oeuvre. (2) Le fait que les répondants cessent de répondre par la poste et qu'ils passent à la méthode de l'IDTC semble

2.5 Analyse des coûts

Après que les essais eurent démontré que tant la méthode de l'ITAO que celle de l'IDTC offraient un rendement élevé et étaient bien acceptées par les répondants, la dernière phase des travaux de recherche a porté sur une analyse des coûts de transition afférents à la méthode de l'ITAO et des coûts courants de mise en oeuvre de l'IDTC.

A cette fin, nous avons étudié les principales catégories de dépenses "salariales" et "non salariales" relatives aux méthodes de collecte par la poste, par ITAO et par IDTC (figure 2). L'étude a porté non seulement sur des estimations des dépenses actuelles, mais aussi sur les dépenses prévues au cours des 10 prochaines années, compte tenu du taux actuel d'accroissement des principaux postes de dépense. Comme la méthode de l'ITAO ne devait être utilisée que de façon transitoire pendant une période de 6 mois (c.-à-d. le temps nécessaire pour amener les unités ayant donné une réponse tardive suivant la méthode de collecte par la poste à répondre dans les délais fixés suivant la méthode de l'ITAO) avant que l'on applique pour de bon la méthode de l'IDTC, l'analyse des coûts a surtout porté sur une comparaison des coûts afférents à la mise en oeuvre de la méthode de l'IDTC.

Sur le plan des catégories de dépenses salariales, l'utilisation de l'IDTC a permis de remplacer les opérations mensuelles d'envoi postal, de retour par la poste, d'enregistrement et de contrôle des formules par une seule opération annuelle d'envoi postal, et donc d'éliminer l'exécution mensuelle d'une importante opération de bureau dans les Etats. De même, les opérations de saisie et de validation de la saisie par lots ainsi que de contrôle des formules qu'impose la méthode de collecte par la poste ont été complètement éliminées grâce à l'IDTC, dans le cadre de laquelle le répondant introduit à l'aide d'un téléphone à clavier les données relatives à son entreprise, puis valide lui-même chacune des entrées. Autre caractéristique d'importance,

Catégorie de dépenses	Poste	ITAO	Auto-interview (IDTC et RP)
DÉPENSES SALARIALES	↙		
Envoi postal	↙		
Retour par la poste	↙		
Saisie des données	↙	↙	
Contrôle et conciliation	↙	↙	↙
Suivi des cas de non-réponse			↙
DÉPENSES NON SALARIALES	↙		
Frais d'affranchissement			↙
Frais d'appel		↙	↙
Micro-ordinateurs		↙	↙

Facteurs récents de variation annuelle des prix

Main-d'oeuvre	+5.7%	ISE, Administrations des Etats et administrations régionales
Affranchissement	+5.0%	Service des postes des E.-U.
Frais d'appel	-1.7%	IPC-E.-U., appels interurbains à l'intérieur d'un même Etat
Micro-ordinateurs	-19.5%	IPA Indices de prix expérimentaux (ordinateur 16 bits)

Figure 2. Coûts de collecte des données (les flèches indiquent la direction des récentes variations de prix)



La durée moyenne d'une interview téléphonique assistée par ordinateur est fonction du nombre d'éléments d'information à recueillir, du rendement temporel de l'appareil utilisé et de l'expérience de la personne recueillant les données. Ainsi, on a pu réduire d'un tiers la durée moyenne des appels d'ITAO (figure 1) suite à la rationalisation du matériel utilisé et à l'accroissement de l'expérience des intervieweurs. Une autre question très importante prise en considération au cours des essais a été l'incidence de la méthode de l'ITAO sur la perte d'effectifs de l'échantillon. On craignait que les employeurs cessent de participer au programme pour éviter d'être constamment ennuyés par des appels téléphoniques. Toutefois, il s'est avéré que le taux de perte d'effectifs de l'échantillon était environ trois fois moins élevé avec la méthode de l'ITAO qu'avec la méthode de collecte par la poste, et qu'avec la première méthode il n'y avait presque pas de répondants d'importance qui sortaient de l'échantillon. Bref, il semble que l'ITAO a reçu un accueil favorable auprès de la majeure partie des répondants et a presque permis d'optimiser le taux de réponse pouvant être obtenu pour les estimations provisoires.

Par suite de l'accroissement des coûts reliés à l'utilisation de la méthode de l'ITAO, le Bureau a entrepris des travaux de recherche sur la collecte des données au moyen d'un téléphone à clavier. Pendant les 4 années qu'ont duré les essais, il s'est révélé possible d'amener les répondants rapides affichant un taux de réponse de 82 à 84% avec la méthode de l'ITAO à afficher des taux de réponse aussi élevés en ayant recours à une méthode de déclaration par IDTC complètement automatisée (figure 1). L'importance de ce résultat s'explique par les économies que l'IDTC permet de réaliser par rapport à l'ITAO. Un des principaux sujets de préoccupation relatifs à la collecte par IDTC était que, contrairement à ce qui se produit pour la collecte par ITAO, où les heures fixées pour les appels s'échelonnent sur toute la journée, il se pourrait que les répondants par IDTC aient tendance à concentrer leurs appels au cours de la même période, entraînant du même coup la transmission de signaux d'occupation et nécessitant l'utilisation d'un nombre excessif de micro-ordinateurs à entrée par téléphone à clavier pour traiter les appels reçus en période de pointe. Heureusement, ces inquiétudes se sont avérées sans fondement et, bien que les micro-ordinateurs à entrée par téléphone à clavier soient en service 24 heures par jour, la majeure partie des appels sont répartis assez uniformément entre 8 h et 17 h (figure 4). Par ailleurs, le pourcentage de répondants auxquels il faut téléphoner pour leur rappeler qu'ils doivent communiquer leurs données a tendance à s'établir au même niveau (environ 40%), qu'on utilise la méthode de l'IDTC ou la méthode de l'ITAO. On met actuellement à l'essai des méthodes visant à réduire le nombre d'appels servant à rappeler aux répondants qu'ils doivent communiquer leurs données nécessités par la méthode de l'IDTC. Un des principaux avantages que présente cette méthode pour le répondant est qu'une interview par IDTC ne prend que la moitié du temps d'une interview réalisée par ITAO, la durée moyenne d'une interview par IDTC s'établissant à seulement 1 minute 45 secondes. En outre, la plupart des entreprises sont déjà pourvues de téléphones à clavier et les estimations courantes indiquent que plus de 80% des employeurs sont en mesure de déclarer les données au moyen de la méthode de l'IDTC. Bien que cette méthode offre de nombreux avantages à l'organisme statistique, sa caractéristique la plus attrayante est la faveur dont elle jouit auprès des répondants. Sa rapidité et sa commodité pour les répondants font que ces derniers réagissent très positivement à cette méthode.

Mentionnons une observation générale relative à l'élaboration d'un programme de recherche sur l'ITAO, soit le fait que la nature du matériel ou du logiciel utilisé au cours des travaux de recherche n'est pas critique, pour autant que ce matériel ou ce logiciel soit assez souple pour être modifié. Il est possible que les résultats finals des essais indiquent que la mise en oeuvre du système d'ITAO comporte des exigences très différentes de celles posées initialement pour le programme de recherche. À cet égard, l'activité la plus importante et celle à laquelle il faut consacrer le plus de temps est l'élaboration et l'amélioration des méthodes et procédures de communication avec les répondants. Une fois qu'on a élaboré des méthodes et procédures efficaces, les exigences posées par la mise en oeuvre du "bon" système deviennent beaucoup plus claires.



2.4 Résultats des travaux de recherche

Au cours des 7 dernières années, le Bureau a été en mesure de déterminer que des données sur la paie sont disponibles dans la plupart des entreprises avant la date limite pour la diffusion des estimations provisoires et que la méthode de collecte par ITAO permet, à l'intérieur d'un délai de 6 mois, d'amener les entreprises ayant traditionnellement donné des réponses tardives (c.-à-d. affichant un taux de réponse de 0 à 20% pour les estimations provisoires) à transmettre leurs réponses dans les délais prévus avec un taux de réponse de 82 à 84% (figure 1). Les taux de réponse ainsi obtenus sont demeurés remarquablement stables au fil des ans, au fur et à mesure que la taille de l'échantillon utilisé pour l'ITAO s'élargissait, pour passer de 400 à 5,000 unités, et que le nombre d'États participants s'accroissait de 2 à 14. Selon les résultats de la recherche, la majorité des entreprises disposent des données à temps pour respecter la date limite pour la diffusion, et l'utilisation de la méthode de l'ITAO permet d'augmenter dans une mesure de 60 à 80% le taux de réponse des entreprises donnant une réponse tardive par la poste, ainsi que de maintenir ce taux à l'intérieur de la plage visée de 80 à 90% sur de longues périodes de temps. On a découvert que le principal facteur limitant la capacité du répondant à respecter la date limite pour la diffusion était la longueur de la période de paie de l'entreprise (figure 3). Il y a généralement une période de paye par semaine, aux deux semaines, deux fois par mois, ou chaque mois. Les données sur les périodes de paie hebdomadaires et bimensuelles peuvent presque toujours être recueillies à temps pour être diffusées, tandis que les données sur les périodes de paie aux deux semaines sont la plupart du temps disponibles dans les délais fixés; toutefois, la majorité des périodes de paie mensuelles se terminent longtemps après la date limite pour la diffusion. L'existence de périodes de paie mensuelles a été un des principaux facteurs limitant les taux de réponse obtenus à l'aide de la méthode de l'ITAO à une plage maximale de 82 à 84%.

Les travaux de recherche sur l'ITAO ont permis d'obtenir plusieurs autres résultats importants. Lorsqu'on utilise cette méthode, environ 60% des répondants sont prêts à transmettre leurs données à la date fixée pour le premier appel, tandis que pour les autres 40% cette première communication téléphonique sert à rappeler à ces personnes qu'elles doivent communiquer leurs données (figure 1). Ce taux est demeuré relativement stable tant d'un État à l'autre que d'une année d'essai à l'autre. On prévoit réaliser un essai à petite échelle afin de déterminer s'il est possible de réduire de façon significative le nombre de rappels nécessaires en faisant parvenir une carte postale de préavis au répondant, quelques jours avant la date fixée pour l'ITAO.

	Poste	ITAO				IDTC et RP			
	Taux de réponse	47%	400	83%	44%	5.6			
	Unités	400	400	84%	42%	5.6			
	Taux de réponse	84%	2000	82%	40%	4.8	45%	78%	45%
	Unités	3000	5000	83%	42%	4.4	600	400	600
	Taux de réponse	84%	5000	82%	41%	3.5	2000	400	2000
	Unités	5000	5000	84%	41%	3.8	5000	400	5000
	Taux de réponse	82%							
	Unités								
	Taux de réponse	84%							
	Unités								
	Taux de réponse	47%							
	Unités	1984	1985	1986	1987	1988	1989	1990	

Figure 1. Sommaire des résultats de la recherche

de la Californie à Berkeley qui a été retenu aux fins de cet essai et utilisé subséquentement pendant toute la durée des travaux de recherche. Après avoir sélectionné un échantillon aléatoire de 200 unités dans chaque Etat, le Bureau a graduellement amélioré les procédures et les systèmes de collecte au cours des 7 années suivantes. Les premiers essais ayant permis d'obtenir des taux de réponse très élevés, ils ont été étendus à 9 Etats en 1986, puis à un total de 14 Etats en 1988. La composition de l'échantillon d'essai a également été modifiée en 1986. Plutôt que de porter sur des échantillons aléatoires sélectionnés au sein de l'échantillon global de la CESS, les essais subséquents ont utilisé uniquement des échantillons aléatoires d'entreprises donnant habituellement une réponse tardive (c.-à-d. d'unités affichant un taux de réponse inférieur à 20% au moment de la date limite pour la diffusion des estimations provisoires). Ainsi, le critère choisi pour évaluer l'efficacité des nouvelles méthodes de collecte par ITAO et par IDTC a été leur capacité de faire passer à un taux permanent stable de 80 à 90% le taux de réponse d'échantillons d'unités ayant affiché un taux de réponse de 0 à 20% à la date limite pour la diffusion des estimations provisoires. En 1990, à la conclusion des travaux de recherche sur l'ITAO, le Bureau utilisait cette méthode pour recueillir des données auprès de plus de 5,000 unités chaque mois et avait réalisé au total plus d'un quart de million d'interviews téléphoniques assistées par ordinateur.

Tandis que l'ITAO se révélait très efficace pour accroître les taux de réponse, en 1985, il était également devenu évident que la collecte permanente par ITAO serait plus coûteuse que la méthode existante de collecte par la poste. À ce moment, de nouveaux travaux ont été entrepris pour chercher à réduire le coût de la méthode de l'ITAO tout en maintenant le taux de réponse mensuel au niveau élevé que cette méthode permettait d'atteindre. Bien qu'on ait par la suite réussi à réduire la durée de la période nécessaire pour réaliser une interview téléphonique assistée par ordinateur, c'est l'utilisation d'une nouvelle méthode de déclaration téléphonique articulée sur micro-ordinateur qui devait permettre d'assurer une réduction spectaculaire des coûts de collecte par ITAO.

En 1985, de nombreuses banques américaines utilisaient, aux fins de l'encassement des chèques aux guichets-automates, une version d'un système de vérification par introduction des données au moyen d'un téléphone à clavier. Après avoir déterminé un système de déclaration au moyen d'un téléphone à clavier articulé sur micro-ordinateur se prêtant à la réalisation d'essais dans le cadre d'une enquête, le Bureau a commencé à effectuer, en 1986, un essai de collecte de données à l'aide de cette technique dans deux Etats. Il convient de noter que l'IDTC n'était alors considérée ni comme une technique destinée à remplacer directement la collecte par la poste, ni comme une solution de rechange pour l'ITAO. L'ITAO avait pour objet, par le biais d'un contact personnel et d'un processus de sensibilisation, d'amener les entreprises donnant d'ordinaire une réponse tardive à fournir leur réponse dans les délais prévus, tandis que l'IDTC devait être utilisée auprès des entreprises répondant par ITAO dans les délais prévus afin de maintenir leur taux de réponse au même niveau élevé tout en permettant une réduction substantielle du coût de collecte par unité. Au cours des 5 années où elle a été utilisée, l'IDTC s'est également révélée être une méthode de collecte téléphonique des données très efficace et très fiable. Au moment où les travaux de recherche sur l'IDTC en arrivent également à leur conclusion, plus de 5,000 unités réparties entre 14 Etats continuent de déclarer chaque mois leurs données au moyen de cette technique et le Bureau a recueilli les données pour un total de plus de 100,000 questionnaires à l'aide de cette nouvelle méthode de déclaration automatisée.

Dans la foulée des travaux sur l'IDTC, le Bureau effectue actuellement plusieurs essais à petite échelle d'un nouveau système de déclaration par reconnaissance de la parole. Les résultats préliminaires de ces essais indiquent que le nouveau système permet d'obtenir des taux de réponse mensuels aussi élevés que l'IDTC, mais que la déclaration par RP a l'avantage important de paraître plus naturelle aux répondants et que ces derniers la préfèrent généralement à l'IDTC. Actuellement, le coût d'acquisition du matériel de RP est environ 15 fois celui du matériel d'IDTC; toutefois, d'ici quelques années, à mesure que ce coût diminuera, la déclaration par RP devrait constituer une solution de rechange rentable pour l'IDTC.



2.2 Méthodes de collecte des données

Compte tenu de la date de diffusion des estimations provisoires, les délais de collecte des données de la CESS sont très serrés. Comme le point de référence de la CESS est la période de paie englobant le 12<sup>e</sup> jour du mois, on ne dispose que de 2 1/2 semaines pour recueillir, introduire, contrôler, totaliser, valider et publier les données. Afin de respecter cet échéancier serré, il faut disposer d'une méthode de collecte permettant d'obtenir les données requises dès que l'entreprise en dispose. Les quatre méthodes de collecte des données étudiées sont décrites à tour de rôle ci-après.

**Poste** – Le questionnaire de la CESS est un questionnaire-navette d'une page, expédié par la poste, sur lequel il y a suffisamment d'espace afin que l'employeur puisse inscrire les données pour 12 mois. Après avoir reçu, chaque mois, le questionnaire par la poste le ou vers le 12<sup>e</sup> jour du mois (c.-à-d. la date de référence de l'enquête), l'employeur y inscrit les éléments d'information requis sur la ligne correspondant au mois en cours. Les cinq éléments d'information de base recueillis sont les suivants: nombre total d'employés, nombre d'employés de sexe féminin, nombre d'employés de production (ou d'employés qui ne sont pas des cadres), heures et gains. Une fois le questionnaire rempli, l'employeur le renvoie par la poste à l'organisme d'Etat, où les données qu'il renferme sont saisies puis contrôlées. Le questionnaire est ensuite classé afin qu'il puisse être réexpédié par la poste à l'employeur pour recueillir les données du rapport du mois suivant. Comme nous l'avons déjà mentionné, ce processus permet actuellement d'obtenir un taux de réponse de 50% au cours des 2 1/2 semaines dont on dispose avant la diffusion des estimations provisoires.

**Interview téléphonique assistée par ordinateur (ITAO)** – Suivant la méthode de l'ITAO, on envoie le questionnaire de la CESS à l'employeur par la poste une seule fois au début de l'année et ce dernier le conserve pour y inscrire les données mensuelles pendant toute l'année. Chaque mois, lorsqu'il dispose des données sur la paie, l'employeur inscrit sur le questionnaire les éléments d'information relatifs à ce mois et attend l'appel, dont la date et l'heure ont déjà été fixées, de l'organisme d'Etat chargé de réaliser l'ITAO. Lorsque l'organisme téléphonique, les données sont recueillies par ITAO et contrôlées, puis on fixe la date et l'heure de l'appel en vue de la collecte des données du mois suivant.

**Introduction des données au moyen d'un téléphone à clavier (IDTC)** – Suivant la méthode de l>IDTC, l'employeur suit les mêmes étapes que pour la méthode de l'ITAO, mais, plutôt que d'attendre l'appel de l'organisme d'Etat chargé de réaliser l'ITAO, il compose lui-même un numéro 800 qui permet de communiquer avec le micro-ordinateur à entrée par téléphone à clavier de cet organisme, puis il se sert du clavier de son téléphone pour introduire les éléments d'information à la suite des messages de guidage appropriés prévus dans l'interview automatisée pour la CESS. Après avoir été introduit, chaque élément d'information est relu pour permettre au répondant de le vérifier.

**Reconnaissance de la parole (RP)** – La déclaration des données par la méthode de reconnaissance de la parole et la collecte des données au moyen d'un téléphone à clavier s'effectuent de la même façon, à la différence près que la première méthode ne nécessite pas l'utilisation d'un téléphone à clavier. Il suffit en effet à l'employeur de lire les données inscrites sur la formule pour que le micro-ordinateur à entrée vocale les traduise et les lui relise pour lui permettre de les vérifier. Le système de RP est un système multilocuteur qui permet la reconnaissance de la parole continue et peut reconnaître les chiffres de 0 à 9 ainsi que les termes "yes" et "no".

2.3 Essais réalisés dans le cadre du programme de recherche

En 1983, le Bureau a entrepris l'élaboration d'un système d'ITAO articulé sur micro-ordinateur destiné à être utilisé dans le cadre d'un essai portant sur deux Etats, dont la mise en oeuvre a commencé en 1984 (figure 1). C'est le système d'ITAO élaboré par l'Université



à cet effet. Au cours des années quatre-vingt, les utilisateurs sont devenus beaucoup plus conscients de l'importance de la qualité des données et, sans que la qualité des produits de la CESS ne se soit nécessairement détériorée, leurs attentes sur le plan de la qualité et de l'aptitude à l'usage des données ont considérablement augmenté. L'émergence de cette nouvelle façon de voir est dans une large mesure attribuable aux travaux de Deming, de Juran et d'autres auteurs portant sur la gestion de la qualité. Pendant les années quatre-vingt on s'est également beaucoup plus concentré sur l'importance et sur les utilisations des statistiques sur la paie de la CESS pour l'établissement d'un diagnostic actuel de l'état de santé de l'économie américaine; toutefois, cet accroissement de l'utilisation et ce gain de visibilité des statistiques de la CESS se sont accompagnés d'un accroissement correspondant de la frustration manifestée par les utilisateurs à l'égard des révisions mensuelles. Nous avons également assisté au cours des années quatre-vingt à de spectaculaires percées technologiques et, en particulier, dans le domaine de la micro-informatique. Cette nouvelle technologie a offert aux organismes statistiques de nombreuses occasions nouvelles d'améliorer le contrôle et la qualité de la collecte des données, notamment: l'interview téléphonique assistée par ordinateur, l'introduction des données au moyen d'un téléphone à clavier, la reconnaissance de la parole, l'auto-interview assistée par ordinateur et le télécopieur. Plusieurs de ces techniques allaient ultimement permettre d'améliorer de façon sensible l'actualité et la qualité des données tout en maintenant les coûts des programmes permanents au même niveau, voire même en les réduisant.

Après avoir réalisé, au cours des années quatre-vingt, des travaux de recherche expérimentale dans le cadre de la CESS et avoir soumis à des essais réels certaines des techniques les plus évoluées alors disponibles de collecte automatisée des données, le Bureau prévoit procéder à une application à grande échelle de ces techniques en 1991.

## 2. PROGRAMME DE RECHERCHE DANS LE CADRE DE LA CESS

### 2.1 Objectifs des travaux de recherche

Au début des années quatre-vingt, le Bureau a effectué pendant sept ans d'importants travaux de recherche sur les causes des réponses tardives et sur les méthodes de collecte susceptibles de permettre un accroissement important des taux de réponse en vue d'améliorer l'établissement des estimations provisoires. Ces travaux visaient à obtenir des réponses aux trois questions de base suivantes:

- L'entreprise dispose-t-elle des données à temps pour pouvoir les transmettre avant la date limite pour la diffusion des estimations préliminaires?
- Existe-t-il des méthodes de collecte des données pouvant permettre l'obtention d'un taux de réponse de 80 à 90% en respectant des contraintes de temps aussi serrées?
- Est-il possible de maintenir le coût de ces méthodes de collecte des données à un niveau à peu près équivalent au coût des actuelles méthodes de collecte par la poste?

Le programme de recherche a permis de mettre au point une méthode de collecte mixte par ITAO/IDTC assurant l'obtention du taux de réponse désiré et le respect des contraintes financières imposées. On trouve dans les sections ci-après une brève description de ces méthodes de collecte assistée par micro-ordinateur et des essais réalisés au cours des travaux de recherche ainsi qu'un exposé des taux de réponse obtenus et une analyse des coûts. Les mémoires de recherche dont il est fait état dans la bibliographie renferment de plus amples renseignements sur ces essais. De plus, des résultats récents sur l'erreur de mesure dans le cas de la collecte effectuée à l'aide de téléphones à clavier sont donnés dans un article de Phipps et Tupek, dans la présente publication.

## 1.2 Current Employment Statistics Survey

Les données sont recueillies par la poste depuis la mise en oeuvre initiale de l'enquête au début du XX<sup>e</sup> siècle. L'utilisation de cette méthode de collecte se traduit par la publication initiale, pour les agrégats de niveau élevé, d'estimations "provisaires" basées seulement sur l'échantillon des questionnaires reçus, puis par la diffusion, deux mois plus tard, d'estimations "définitives" portant sur l'ensemble de l'échantillon. L'établissement d'estimations provisoires et définitives pour un mois donne nécessairement qu'on procède périodiquement à une révision substantielle des estimations initiales. Ces révisions ont une incidence non seulement sur les statistiques de base de la CESS, mais aussi sur les autres statistiques utilisant les estimations de la CESS comme intrant. Pour surmonter cette difficulté, le Bureau a mis en oeuvre un programme de recherche sur les techniques de collecte téléphonique automatisée des données dans le but de réduire considérablement l'ampleur et la fréquence des grosses révisions apportées aux estimations provisoires. Le présent article donne un aperçu des travaux de recherche effectués par le Bureau pendant 7 ans sur les techniques de collecte téléphonique automatisée des données et résume certains des principaux résultats de ces travaux. Vous trouverez dans les sections ci-après une description du processus de mise en oeuvre de la CESS, une analyse du programme de recherche évaluant les méthodes de collecte des données par interview téléphonique assistée par ordinateur (ITAO), par introduction des données au moyen d'un téléphone à clavier (IDTC) et par reconnaissance de la parole (RP), un exposé détaillé de certains des principaux résultats des travaux de recherche ayant trait au rendement et au coût de ces méthodes, enfin, pour conclure, une analyse du programme d'application à grande échelle de ces techniques dans le cadre de la CESS.

Avec son échantillon de 350,000 unités, la CESS est la plus importante enquête-échantillon mensuelle effectuée aux États-Unis. Cette enquête est réalisée par le Bureau dans le cadre d'un programme de coopération entre le gouvernement fédéral et les gouvernements des États aux termes duquel il revient au Bureau de préciser le plan de sondage et les procédures opérationnelles de l'enquête, tandis qu'il incombe à chaque État de mettre en oeuvre les activités de collecte des données ainsi que les activités de conciliation. Le Bureau produit et publie, pour l'ensemble du pays, des estimations mensuelles complètes pour toutes les industries au niveau des catégories à 2, 3 ou 4 chiffres, tandis que chaque État produit des estimations mensuelles à l'échelle de l'État et de la région (270 régions statistiques métropolitaines).

à l'échelle de l'Etat et de la région (270 régions statistiques métropolitaines). On s'entend pour reconnaître que les estimations de la CESS constituent des statistiques économiques très précises. Une fois l'an, on obtient à partir des dossiers d'impôt de l'assurance-chômage les chiffres de l'emploi complets (ou pour l'ensemble de l'univers) pour l'année précédente et on utilise ces chiffres pour procéder à un étalonnage (réalignement) annuel des estimations produites à partir de l'échantillon de la CESS en fonction de ces chiffres obtenus pour l'univers. Cet étalonnage annuel permet d'obtenir des estimations plus précises pour le mois en cours ainsi qu'une estimation annuelle de l'erreur globale affectant l'enquête. Au cours des cinq dernières années, l'écart moyen entre l'estimation définitive fondée sur l'échantillon de la CESS et le chiffre obtenu pour l'ensemble de l'univers a été inférieur à 0,2%, et il a été pratiquement nul pendant 4 ans au cours des années quatre-vingt. Alors que l'on considère que les estimations mensuelles définitives de la CESS correspondent très étroitement aux chiffres obtenus pour l'ensemble de l'univers, les estimations mensuelles provisoires, fondées sur environ 50% des questionnaires de l'échantillon reçus par la poste, ont fait périodiquement l'objet d'importantes révisions quand on les a comparées aux estimations définitives publiées deux mois plus tard. Bien qu'on ait pu, au fil des ans, apporter des améliorations permettant de réduire l'ampleur des révisions mensuelles, on considère l'exécution d'importantes révisions périodiques comme un sous-produit de la mise en oeuvre d'une enquête décentralisée de grande envergure utilisant la méthode de collecte de données par la poste.

Un certain nombre de changements survenus au cours de la dernière décennie devaient avoir une incidence considérable sur le programme de la CESS, en influant tant sur l'urgence de résoudre la question des révisions mensuelles que sur les options susceptibles d'être choisies



# Amélioration de la qualité des données à l'aide d'un mode de collecte mixte

GEORGE S. WERKING et RICHARD L. CLAYTON<sup>1</sup>

## RÉSUMÉ

L'établissement d'estimations assujetties à des contraintes de temps pose un problème constant, qui est celui des limitations importantes qui se rattachent à la collecte des données par la poste. Pour surmonter cette difficulté, le Bureau of Labor Statistics des États-Unis, a effectué pendant 7 ans d'importants travaux de recherche sur l'utilisation de l'interview téléphonique assistée par ordinateur (ITAO) et de l'auto-interview assistée par ordinateur (AIAO), dans le cadre de cette dernière méthode les répondants introduisant eux-mêmes leurs données au moyen d'un téléphone à clavier ou par reconnaissance de la parole. Le présent article donne un aperçu de certains des principaux résultats de ces travaux ayant trait au rendement et au coût de ces méthodes. L'exposé s'achève par des observations sur un programme d'application à grande échelle de ces techniques à un échantillon mensuel de 350,000 entreprises.

**MOTS CLÉS:** Statistiques sur l'emploi; révisions; ITAO; collecte des données au moyen d'un téléphone à clavier; collecte des données par reconnaissance de la parole; analyse des coûts.

## 1. INTRODUCTION

### 1.1 Les statistiques sur l'emploi aux États-Unis

Le premier vendredi de chaque mois, le Bureau of Labor Statistics des États-Unis diffuse des données sur la situation de l'emploi aux États-Unis pour le mois précédent. Le jour de la publication des données, le Commissionner of Labor Statistics se présente devant le Joint Economic Committee du Congrès pour faire une analyse détaillée des données et des tendances relatives au mois en cours; au même moment, les données sont mises à la disposition des médias d'information ainsi que de la communauté financière et du monde des affaires. Cet ensemble de statistiques, qui est suivi de près par les observateurs, constitue le premier indicateur de l'activité économique pour le mois précédent et il est utilisé comme un des principaux étalons de mesure de la santé de l'économie américaine. Les statistiques publiées comprennent des données sur l'emploi, sur les gains et sur les heures par industrie, au niveau détaillé, recueillies au moyen de l'enquête mensuelle réalisée par le Bureau auprès de 350,000 entreprises – la Current Employment Statistics Survey (CESS) – ainsi que de données sur la population active et sur le chômage, recueillies au moyen d'une enquête réalisée par le Bureau auprès d'un échantillon de 60,000 ménages – la Current Population Survey (CPS). Les données de l'enquête sur les entreprises trouvent nombre d'utilisations importantes sur le plan économique. Étant donné l'envergure et l'actualité de la CESS ainsi que l'importance des statistiques de base sur la paie qu'elle permet de recueillir, les estimations mensuelles de l'enquête sont non seulement utilisées toutes seules comme indicateurs économiques principaux, mais elles servent également à l'élaboration de nombreux autres indicateurs importants de l'activité économique du pays, y compris: le revenu personnel pour le calcul du produit national brut, l'indice des indicateurs économiques avancés, l'indice des indicateurs instantanés d'activité, l'indice de l'activité industrielle, certaines mesures des gains réels et certaines mesures de la productivité. Bien que l'actualité et l'exactitude des statistiques de la CESS soient des caractéristiques essentielles pour l'analyse des conditions économiques qui prévalent aux États-Unis,

<sup>1</sup> G. Werking et R. Clayton, Monthly Industry Employment Statistics Division, Bureau of Labor Statistics, Bureau 2089, 441 G Street, N.W., Washington, D.C., E.-U., 20212.



Kott montre comment certaines techniques élaborées en vertu de la théorie des sondages fondée sur les plans, notamment l'inclusion des poids d'échantillonnage et l'estimateur de l'erreur quadratique moyenne fondé sur la linéarisation, peuvent servir à l'estimation d'un système d'équations linéaires. Il montre aussi que l'utilisation de poids d'échantillonnage peut être souhaitable en l'absence probable de variables explicatives. De plus, l'estimateur de l'erreur quadratique moyenne fondé sur la linéarisation est quasi-non biaisé pour de nombreuses structures d'erreur.

La nécessité de publier le plus rapidement possible et, en même temps, de pouvoir établir des estimations justes à partir de toutes les données disponibles amène l'élaboration d'un processus de révision des estimations des comptes nationaux. Biggeri et Trivellato examinent ce qui s'est fait récemment au chapitre de l'analyse de la fiabilité des estimations des comptes nationaux révisées ultérieurement. Ils présentent aussi une étude empirique où ils utilisent des données du Canada, de l'Italie et des États-Unis.

Le rédacteur en chef

Au moment de mettre sous presse, nous apprenons le décès de M. M.N. Murthy, Directeur de l'Applied Statistics Centre, Madras, Inde, le 2 avril, 1991. M. Murthy a fait des contributions importantes à la méthodologie d'enquête. Il était l'auteur du livre très reconnu *Sampling Theory and Methods*. Pendant sa carrière, il a travaillé pour l'Indian Statistical Institute et l'Institut Statistique de l'ONU pour l'Asie et le Pacifique. Il était Fellow de l'ASA, la Royal Statistical Society et l'International Statistical Institute. Nous étions comblés de l'avoir comme membre du comité de rédaction. Il manquera à ses collègues et ses étudiants de par le monde. Il laisse dans le deuil sa femme, Vyjayanthi, et sa fille Shashi.

## Dans ce numéro

La réalisation d'enquêtes téléphoniques par les méthodes classiques est relativement peu coûteuse et assure un contact direct, bien qu'à distance, entre l'intervieweur et le répondant. Ce sont surtout ces deux caractéristiques – coût peu élevé et contact humain – qui ont fait de l'enquête téléphonique un outil de sondage privilégié. Les nombreuses recherches qui ont été faites à ce sujet durant la dernière décennie sont résumées dans *Telephone Survey Methodology*, compilé par R.M. Groves et coll. (1988), et dans le *Journal of Official Statistics*, vol. 4, n° 4. La section spéciale de ce numéro, consacrée à la collecte et à la saisie des données, fait état d'autres recherches où le téléphone joue encore un rôle important dans un contexte technologique plus nouveau.

Dans le premier article de la section spéciale, Werking et Clayton commentent les recherches qu'a réalisées depuis sept ans le U.S. Bureau of Labor Statistics (BLS) sur les méthodes de collecte de données par téléphone. Ils montrent comment ces recherches ont conduit à la décision de mettre en oeuvre l'interview téléphonique assistée par ordinateur (ITAO) et la saisie de données au moyen d'un téléphone à clavier (TDE – touchtone data entry) dans la Current Employment Statistics Survey en 1991 et 1992. Les auteurs soulignent aussi une nouvelle technique, la reconnaissance de la parole, qui peut remplacer la saisie par boutons-poussoirs.

Phlips et Tupek examinent un des aspects des enquêtes téléphoniques sur lesquels ont porté les recherches du BLS, notamment les erreurs de mesure dans les enquêtes avec saisie au moyen d'un téléphone à clavier. Ils concluent que la plus grande source d'erreurs est l'introduction de chiffres en trop ou de chiffres erronés mais que cela peut être corrigé partiellement grâce au contrôle des données, surtout lorsqu'il existe des données longitudinales. Ils proposent aussi des moyens d'améliorer les systèmes de saisie au moyen d'un téléphone à clavier.

La composition aléatoire (CA) de Mitofsky-Waksberg est une méthode courante de sélection des ménages. Brick et Waksberg examinent une version modifiée dont ils étudient les propriétés statistiques. De cette méthode et proposent une version modifiée dont ils étudient les propriétés statistiques. De plus, ils indiquent des critères permettant de déterminer dans quelles circonstances on doit opter pour la méthode originale ou ses variantes.

Bethlehem et Keller présentent le système Blaise, un logiciel innovateur élaboré au Bureau de la statistique des Pays-Bas. Les auteurs montrent comment Blaise sert à intégrer les diverses étapes du traitement des données d'enquête, y compris la collecte et la saisie. Avec l'utilisation croissante des ordinateurs portatifs dans les enquêtes par sondage, des outils comme le système Blaise deviendront indispensables.

Comme la plupart des organismes de statistiques à l'étranger, Statistique Canada étudie la possibilité d'étendre et de perfectionner l'utilisation du téléphone dans les enquêtes par sondage. Dans le dernier article de la section spéciale, Drew résume les recherches et les expériences récentes qui ont été faites en ce qui concerne les enquêtes-ménages. Il expose aussi les conséquences de ces recherches et de ces expériences pour le remaniement de l'Enquête sur la population active.

On peut utiliser un modèle de régression linéaire normal pour l'échantillonnage répété. Bellhouse obtient les fonctions de vraisemblance marginales et conditionnelles pour la matrice des coefficients de corrélation de ce modèle. Il présente des applications de ces fonctions pour l'échantillonnage aléatoire simple et pour des plans plus complexes.

Schiopu-Kratina et Srinath décrivent la méthodologie de l'Enquête sur l'emploi, la rémunération et les heures de travail (EBRH) de Statistique Canada. L'EBRH est une grande enquête mensuelle menée auprès des entreprises et qui repose sur un échantillon avec renouvellement. La détermination de la taille de l'échantillon à chaque mois est une opération assez complexe vu l'évolution constante de la population. Les auteurs proposent des façons de simplifier le plan de sondage.





# TECHNIQUES D'ENQUÊTE

Une revue de Statistique Canada  
Volume 17, numéro 1, juin 1991

## TABLE DES MATIÈRES

Dans ce numéro .....	1
Nouvelles approches pour la collecte et la saisie des données	
G.S. WERKING et R.L. CLAYTON	
Amélioration de la qualité des données à l'aide d'un mode de collecte mixte .....	3
P.A. PHIPPS et A.R. TUPEK	
Fiabilité des données introduites au moyen d'un téléphone à clavier .....	17
J.M. BRICK et J. WAKSBERG	
Méthode pour éviter l'échantillonnage progressif dans une enquête téléphonique à composition aléatoire .....	31
J.G. BETHLEHEM et W.J. KELLER	
Le système Blaise ou comment élaborer un système de traitement intégré des données d'enquête .....	47
J.D. DREW	
Recherche et essais pour les méthodes d'enquêtes par téléphone à Statistique Canada .....	63
D.R. BELLHOUSE	
Fonctions de vraisemblance marginales et fonctions de vraisemblance conditionnelles approximatives pour l'échantillonnage répété .....	77
I. SCHIOPU-KRATINA et K.P. SRINATH	
Renouvellement de l'échantillon et estimation dans l'Enquête sur l'emploi, la rémunération et les heures de travail .....	89
P.S. KOTT	
Estimation d'un système d'équations linéaires à l'aide de données d'enquête .....	101
L. BIGGERI et U. TRIVELLATO	
L'évaluation des erreurs dans les données de comptabilité nationale: les estimations provisoires et révisées .....	111

# TECHNIQUES D'ENQUÊTE

## Une revue de Statistique Canada

La revue Techniques d'enquête est répertoriée dans The Survey Statistician et Statistical Theory and Methods Abstracts. On peut en trouver les références dans Current Index to Statistics, et Journal Contents in Qualitative Methods.

### COMITÉ DE DIRECTION

#### Président

G.J. Brackstone  
B.N. Chinnappa

#### Membres

G.J.C. Hole  
C. Patrick  
F. Mayda (Directeur de la production)  
R. Platak  
D. Roy  
M.P. Singh

### COMITÉ DE RÉDACTION

#### Rédacteur en chef

M.P. Singh, *Statistique Canada*

#### Rédacteurs associés

B. Afonja, *Nations Unies*

D.R. Bellhouse, *U. of Western Ontario*

D. Binder, *Statistique Canada*

E.B. Dagum, *Statistique Canada*

J.-C. Deville, *INSEE*

D. Drew, *Statistique Canada*

W.A. Fuller, *Iowa State University*

J.F. Gentleman, *Statistique Canada*

M. Gonzalez, *U.S. Office of Management and Budget*

R.M. Groves, *U.S. Bureau of the Census*

#### Rédacteurs adjoints

J. Gambino, L. Mach et A. Thêberge, *Statistique Canada*

### POLITIQUE DE RÉDACTION

La revue Techniques d'enquête publie des articles sur les divers aspects des méthodes statistiques qui intéressent un organisme statistique comme, par exemple, les problèmes de conception découlant de contraintes d'ordre pratique, l'utilisation de différentes sources de données et de méthodes de collecte, les erreurs dans les enquêtes, l'évaluation des enquêtes, la recherche sur les méthodes d'enquête, l'analyse des séries chronologiques, la désaisonnalisation, les études démographiques, l'intégration de données statistiques, les méthodes d'estimation et d'analyse de données et le développement de systèmes généralisés. Une importance particulière est accordée à l'élaboration et à l'évaluation de méthodes qui ont été utilisées pour la collecte de données ou appliquées à des données réelles. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de Statistique Canada.

### Présentation de textes pour la revue

La revue Techniques d'enquête est publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à en faire parvenir le texte au rédacteur en chef, M. M. P. Singh, Division des méthodes d'enquêtes sociales, Statistique Canada, Tunney's Pasture, Ottawa (Ontario), Canada K1A 0T6. Prière d'envoyer quatre exemplaires dactylographiés selon les directives présentées dans la revue. Ces exemplaires ne seront pas retournés à l'auteur.

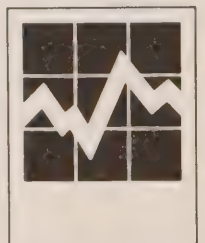
### Abonnement

Le prix de la revue Techniques d'enquête (catalogue n° 12-001) est de 35 \$ par année au Canada, 42 \$ (É.-U.) aux États-Unis, et de 49 \$ (É.-U.) par année à l'étranger. Prière de faire parvenir votre demande d'abonnement à Section des ventes des publications, Statistique Canada, Ottawa (Ontario), Canada K1A 0T6. Un prix réduit est offert aux membres de l'American Statistical Association, l'Association Internationale de Statisticiens d'Enquête et la Société Statistique du Canada.

Statistique Canada  
Division des méthodes d'enquêtes sociales

# Techniques d'enquête

Une revue de Statistique Canada  
Juin 1991 Volume 17 Numéro 1



Publication autorisée par le ministre de  
l'Industrie, des Sciences et de la Technologie

© Ministre des Approvisionnements  
et Services Canada 1991

Tous droits réservés. Il est interdit de reproduire ou de  
transmettre le contenu de la présente publication, sous quelque  
forme ou par quelque moyen que ce soit, enregistré ou sur  
support magnétique, reproduction électronique, mécanique,  
photographique, ou autre, ou de l'immaginer dans un système  
de recouvrement, sans l'autorisation écrite préalable du ministre  
des Approvisionnements et Services Canada.

Juin 1991

Canada : 35 \$

États-Unis : 42 \$ US

Autres pays : 49 \$ US

Catalogue 12-001

ISSN 0714-0045

Ottawa







# Techniques d'enquête

Une revue de Statistique Canada

Juin 1991 / Volume 17 Numéro 1



12  
-001

Catalogue 12-001

# Survey Methodology

A Journal of Statistics Canada

December 1991      Volume 17    Number 2



Statistics  
Canada

Statistique  
Canada

Canada







Statistics Canada  
Social Survey Methods Division

# Survey Methodology

A Journal of Statistics Canada

December 1991      Volume 17   Number 2

Published by authority of the Minister  
responsible for Statistics Canada

© Minister of Industry,  
Science and Technology, 1991

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise without prior written permission from Chief, Author Services, Publications Division, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6.

December 1991

Price: Canada: \$35.00

United States: US\$42.00

Other Countries: US\$49.00

Catalogue 12-001

ISSN 0714-0045

Ottawa

# SURVEY METHODOLOGY

## A Journal of Statistics Canada

The Survey Methodology Journal is abstracted in The Survey Statistician and Statistical Theory and Methods Abstracts and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods.

### MANAGEMENT BOARD

<b>Chairman</b>	G.J. Brackstone	
<b>Members</b>	B.N. Chinnappa	C. Patrick
	G.J.C. Hole	D. Roy
	F. Mayda (Production Manager)	M.P. Singh
	R. Platek (Past Chairman)	

### EDITORIAL BOARD

**Editor** M.P. Singh, *Statistics Canada*

#### Associate Editors

B. Afonja, <i>United Nations</i>	G. Kalton, <i>University of Michigan</i>
D.R. Bellhouse, <i>U. of Western Ontario</i>	J.N.K. Rao, <i>Carleton University</i>
D. Binder, <i>Statistics Canada</i>	L.-P. Rivest, <i>Laval University</i>
E.B. Dagum, <i>Statistics Canada</i>	D.B. Rubin, <i>Harvard University</i>
J.-C. Deville, <i>INSEE</i>	I. Sande, <i>Bell Communications Research, U.S.A.</i>
D. Drew, <i>Statistics Canada</i>	C.E. Särndal, <i>University of Montreal</i>
W.A. Fuller, <i>Iowa State University</i>	W.L. Schaible, <i>U.S. Bureau of Labor Statistics</i>
J.F. Gentleman, <i>Statistics Canada</i>	F.J. Scheuren, <i>U.S. Internal Revenue Service</i>
M. Gonzalez, <i>U.S. Office of Management and Budget</i>	C.M. Suchindran, <i>University of North Carolina</i>
R.M. Groves, <i>U.S. Bureau of the Census</i>	J. Waksberg, <i>Westat Inc.</i>
D. Holt, <i>University of Southampton</i>	K.M. Wolter, <i>A.C. Nielsen, U.S.A.</i>

#### Assistant Editors

J. Gambino, L. Mach, H. Mantel and A. Thériberge, *Statistics Canada*

---

### EDITORIAL POLICY

The Survey Methodology Journal publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

#### Submission of Manuscripts

The Survey Methodology Journal is published twice a year. Authors are invited to submit their manuscripts in either of the two official languages, English or French to the Editor, Dr. M.P. Singh, Social Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6. Four nonreturnable copies of each manuscript prepared following the guidelines given in the Journal are requested.

#### Subscription Rates

The price of the Survey Methodology Journal (Catalogue No. 12-001) is \$35 per year in Canada, US \$42 in the United States, and US \$49 per year for other countries. Subscription order should be sent to Publication Sales, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6. A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, and the Statistical Society of Canada.



# **SURVEY METHODOLOGY**

A Journal of Statistics Canada  
Volume 17, Number 2, December 1991

## **CONTENTS**

In This Issue .....	121
J.M. ALHO	
Variance Estimation in Dual Registration Under Population Heterogeneity .....	123
P.S.R.S. RAO and I.M. SHIMIZU	
Combining Estimates from Surveys .....	131
P.J. LAVRAKAS, R.A. SETTERSTEN, Jr. and R.A. MAIER, Jr.	
RDD Panel Attrition in Two Local Area Surveys .....	143
B.C. SUTRADHAR, E.B. DAGUM and B. SOLOMON	
An Exact Test for the Presence of Stable Seasonality With Applications .....	153
J.-C. DEVILLE	
A Theory of Quota Surveys .....	163
G. KALTON	
Sampling Flows of Mobile Human Populations .....	183
M.A. HIDIROGLOU, G.H. CHOUDHRY and P. LAVALLÉE	
A Sampling and Estimation Methodology for Sub-Annual Business Surveys .....	195
E.A. STASNY, P.K. GOEL and D.J. RUMSEY	
County Estimates of Wheat Production .....	211
D.A. NORRIS and D.G. PATON	
Canada's General Social Survey: Five Years of Experience .....	227
Acknowledgements .....	241



## In This Issue

This issue of *Survey Methodology* contains papers covering a broad range of topics. In the first paper, Alho investigates different estimators of the variance of the population size estimated using dual registration. The bias of the usual variance estimator, derived under the assumption of homogeneous capture probabilities, is investigated under population heterogeneity and two alternative variance estimators are proposed. The three estimators are applied to occupational disease data from Finland.

Rao and Shimizu compare three procedures for combining independent estimates obtained at successive time periods. They show that all three procedures produce an improvement over the use of the estimate based on only one occasion. The findings are illustrated by applying them to data from the American National Health Discharge Survey.

In their paper, Lavrakas, Settersten and Maier take a descriptive look at the problem of panel attrition in surveys, using data from two surveys carried out using random digit dialing. The paper gives the reader a good introduction to some of the causes of attrition, with suggestions on how its effects can be reduced.

Sutradhar, Dagum and Solomon give an exact test for the presence of significant stable seasonality for time series with seasonal patterns that are stable over time except for possible annual shifts. The assumptions of the standard ANOVA F-test used by the X-11-ARIMA seasonal adjustment method are violated when the residuals are autocorrelated. The exact test, however, takes into account the possibility of autocorrelated residuals. The exact and the standard tests are compared for several Canadian socioeconomic series.

A characteristic of quota sampling is the absence of randomized selection. Some kind of modelling must be imposed to construct estimators. The traditional approach is to use superpopulation modelling, and Deville proposes interesting extensions to this method. An approach suggested by the author is to model the sampling process. Comparisons with random sampling are made.

While household surveys are successfully used to collect data about human populations, they are not suitable for studying the characteristics of mobile human populations, such as visitors to museums or parks, shoppers, *etc.* Kalton describes different sample designs for surveys of flows of human populations and provides a number of examples of such surveys. The examples illustrate that field work considerations play an important role in the choice of a sample design.

Hidioglou, Choudhry and Lavallée provide a sampling methodology for continuing sub-annual business surveys. A rotation scheme is suggested to maintain a representative sample through time. The properties of a number of estimators of totals for this sampling methodology have been evaluated in an empirical study which reflects a number of possible survey conditions.

Stasny, Goel and Rumsey use regression models to obtain small area estimates of wheat production when the data come from non-probabilistic sources. A simulation study compares the estimates obtained through this approach with the standard synthetic and direct estimators. Three scaling methods to satisfy additivity constraints are also compared.

In the last paper of this issue, Norris and Paton give an overview of Canada's five year old General Social Survey. They present a brief account of the information needs and discuss the five annual topics addressed by the survey. A description of the survey's methodology and experiences with the use of random digit dialing are presented. The authors' analysis of nonresponse rates over the life of the survey has implications for other telephone surveys.





## Variance Estimation in Dual Registration Under Population Heterogeneity

JUHA M. ALHO<sup>1</sup>

### ABSTRACT

The usual dual system estimator for population size can be severely biased, if there is population heterogeneity in the capture probabilities. In this note we investigate the bias of the corresponding variance estimator under heterogeneity. We show that the usual estimator is conservative, *i.e.*, it gives too large values, if the two registration systems are negatively correlated, uncorrelated, or when the correlation is positive, but small. In the case of high positive correlation the usual estimator may yield too low values. Two alternative estimators are proposed. One is conservative under arbitrary heterogeneity. The other is conservative under Gaussian heterogeneity. The methods are applied to occupational disease data from Finland.

KEY WORDS: Capture-recapture; Dual system; Heterogeneity; Occupational diseases.

### 1. INTRODUCTION

Suppose there are  $N$  individuals in a closed population. The problem is to estimate the unknown  $N$  using dual registration. We sample twice with  $n_j$  individuals captured at the  $j$ th time,  $j = 1, 2$ . Let  $m$  be the number captured twice. Define indicator variables  $u_{ji}$  and  $m_i$  for  $i = 1, \dots, N$  such that  $u_{ji} = 1$ , if and only if individual  $i$  is captured at the  $j$ th time only,  $j = 1, 2$ ; and  $m_i = 1$ , if and only if individual  $i$  is captured twice. Otherwise  $u_{ji}$  and  $m_i$  are zero. Define  $n_{ji} = u_{ji} + m_i$  as the indicator of capture at the  $j$ th time,  $j = 1, 2$ . Let  $M_i = u_{1i} + u_{2i} + m_i$  indicate capture at least once. Define the individual capture probabilities as  $p_{ji} = E[n_{ji}]$ ,  $j = 1, 2$ ; and  $p_{12i} = E[m_i]$ . Assume that the probabilities are strictly between zero and one. The fact that the probabilities are allowed to vary by individual indicates that we may have population heterogeneity in the capture probabilities. We complete the definition of the dual registration (or capture-recapture) model by assuming that the captures are independent for each individual, or  $p_{12i} = p_{1i}p_{2i}$ , and that the multinomial vectors

$$(u_{1i}, u_{2i}, m_i, 1 - M_i) \sim \text{Mult}(1; p_{1i}q_{2i}, p_{2i}q_{1i}, p_{1i}p_{2i}, 1 - \phi_i),$$

where  $q_{ji} = 1 - p_{ji}$ ,  $j = 1, 2$ , and  $\phi_i = p_{1i} + p_{2i} - p_{1i}p_{2i}$ , are independent for  $i = 1, \dots, N$ .

It is well-known that when capture probabilities do not vary by individual, or  $p_{ji} = p_j$ ,  $j = 1, 2$ , the maximum likelihood estimator of  $N$  is  $\hat{N} = n_1n_2/m$  (or more precisely, the largest integer short of this value; *cf.*, Feller 1968, p. 46). This classical estimator can be severely biased under population heterogeneity (Seber 1982, p. 565; Burnham and Overton 1979, Table 4, pp. 931-932). As shown, *e.g.*, in Example 1 below, under homogeneous capture probabilities the asymptotic variance of  $\hat{N}$  is  $\text{Var}(\hat{N}) = Nq_1q_2/(p_1p_2)$ , where  $q_j = 1 - p_j$ ,  $j = 1, 2$ . Then  $\text{Var}(\hat{N})$  can be estimated by  $V_1 = n_1n_2u_1u_2/m^3$  (Sekar and Deming 1949, pp. 114-115).

<sup>1</sup> Juha M. Alho, Institute for Environmental Studies and Department of Statistics, University of Illinois at Urbana-Champaign, 1101 W. Peabody Dr., Urbana IL, 61801, U.S.A.

The purpose of this note is to investigate the adequacy of the variance estimator  $V_1$ , and compare the bias of  $V_1$  to the bias of  $\hat{N}$ . One motive for investigating  $V_1$  is that it has not been previously known whether  $V_1$  is adequate in the case in which there is population heterogeneity, but  $\hat{N}$  is, nevertheless, consistent. This turns out to be the case. Similarly, it has not been clear when  $V_1$  gives overestimates and thus can lead to valid confidence intervals, despite the bias of  $\hat{N}$ . This turns out to be possible for one-sided intervals in special circumstances.

In Section 2 we calculate the asymptotic variance of  $\hat{N}$ , as  $N \rightarrow \infty$ , and derive a conservative estimator  $V_2$  for this variance under arbitrary heterogeneity. In other words,  $V_2$  overestimates the true asymptotic variance. One might hope that an overestimate of variance could compensate for the typically negative bias of  $\hat{N}$  and still yield valid confidence intervals. Unfortunately, this appears possible only when the bias of  $\hat{N}$  is small, or when  $N$  is small. In Section 3 the adequacy of  $V_1$  is studied under Gaussian heterogeneity and an estimator  $V_3$  is derived, which is conservative under this restricted type of heterogeneity. Gaussianity *per se* is not required for the arguments, only that the moments of the pairs  $(p_{1i}, p_{2i})$  agree with those of a bivariate Gaussian distribution. This setup permits the ready examination of the effect of correlation between  $p_{1i}$ 's and  $p_{2i}$ 's on variance estimation, because correlation is expressible in terms of just one parameter, the ordinary moment correlation coefficient. In Section 4 we compare the bias in variance estimates to the bias of  $\hat{N}$  using empirical data relating to the registration of occupational diseases in Finland.

## 2. BIAS AND VARIANCE UNDER HETEROGENEITY

Define  $\bar{p}_{jN}$  as the average probability of capture at the  $j$ th time,  $j = 1, 2$ ; and let  $\bar{p}_{12N}$  be the average of the products  $p_{1i}p_{2i}$ ,  $i = 1, \dots, N$ . Then,  $C_N = \bar{p}_{12N} - \bar{p}_{1N}\bar{p}_{2N}$  is the covariance of the pairs  $(p_{1i}, p_{2i})$ . Assume that the limits  $\bar{p}_{jN} \rightarrow \bar{p}_j$ ,  $j = 1, 2$ ;  $\bar{p}_{12N} \rightarrow \bar{p}_{12}$ , and  $C_N \rightarrow C$  exist. Then we have that  $\hat{N}/N \rightarrow \bar{p}_1\bar{p}_2/\bar{p}_{12}$ , so  $\hat{N}/N - 1 \rightarrow -C/\bar{p}_{12}$ , as  $N \rightarrow \infty$ . This is the asymptotic bias of the classical estimator under population heterogeneity. Interestingly, it only depends on the first two moments of the distribution of the pairs  $(p_{1i}, p_{2i})$ . As is well-known (Sekar and Deming 1949, pp. 105-106; Seber 1982, p. 86), when the covariance is zero ( $C = 0$ ), then the classical estimator is consistent; if  $C > 0$ ,  $\hat{N}$  gives an underestimate; and if  $C < 0$ , it gives an overestimate. As noted above the adequacy of  $V_1$ , when the  $p_{ji}$ 's vary from one individual to the next but still  $C = 0$ , is of particular interest.

We shall now calculate the asymptotic variance of the classical estimator under our general heterogeneity model. Note that the finite variance does not exist, because there is a positive probability that  $m = 0$ . Therefore, "asymptotic variance" properly refers here to the variance of the limiting distribution rather than to limit of the variances, as  $N \rightarrow \infty$ .

**Lemma 1.** The asymptotic variance of  $\hat{N}$  is

$$\begin{aligned} \text{Var}(\hat{N}) = N \left\{ \frac{\bar{p}_1^2 \bar{p}_2^2}{\bar{p}_{12}^3} - \frac{\bar{p}_1^2 \bar{p}_2}{\bar{p}_{12}^2} - \frac{\bar{p}_1 \bar{p}_2^2}{\bar{p}_{12}^2} - \frac{\bar{p}_2^2}{\bar{p}_{12}^2} \bar{S}_1 - \frac{\bar{p}_1^2}{\bar{p}_{12}^2} \bar{S}_2 \right. \\ \left. - \frac{\bar{p}_1^2 \bar{p}_2^2}{\bar{p}_{12}^4} \bar{S}_3 + 2 \left( \frac{\bar{p}_1 \bar{p}_2^2}{\bar{p}_{12}^3} \bar{S}_4 + \frac{\bar{p}_1^2 \bar{p}_2}{\bar{p}_{12}^3} \bar{S}_5 \right) \right\}, \end{aligned}$$

where  $\bar{S}_j = S_j/N$  for  $j = 1, \dots, 5$ , with



$$S_1 = \sum_{i=1}^N p_{1i}^2, \quad S_2 = \sum_{i=1}^N p_{2i}^2, \quad S_3 = \sum_{i=1}^N p_{1i}^2 p_{2i}^2,$$

$$S_4 = \sum_{i=1}^N p_{1i}^2 p_{2i}, \quad S_5 = \sum_{i=1}^N p_{1i} p_{2i}^2.$$

The proof is sketched in the Appendix. We note that unlike the bias of  $\hat{N}$  that depends on the first two moments of the pairs  $(p_{1i}, p_{2i})$  only,  $\text{Var}(\hat{N})$  depends on moments up to fourth order. In special cases, such as the ones considered in Example 2 and Proposition 2, a simpler representation is possible.

**Example 1.** Suppose there is no heterogeneity in the probabilities, or  $p_{ji} = p_j, j = 1, 2$ . Then  $\bar{p}_j = p_j, j = 1, 2; \bar{p}_{12} = p_1 p_2; \bar{S}_j = p_j^2, j = 1, 2; \bar{S}_3 = p_1^2 p_2^2, \bar{S}_4 = p_1^2 p_2,$  and  $\bar{S}_5 = p_1 p_2^2$ . Hence, the asymptotic variance is  $\text{Var}(\hat{N}) = N(1 - p_1 - p_2 + p_1 p_2)/(p_1 p_2) = Nq_1 q_2/(p_1 p_2)$ . Consistent estimators for  $Np_1 p_2$  and  $Np_j$  are  $m$  and  $n_j, j = 1, 2$ . In other words,  $Np_j/n_j \rightarrow 1, j = 1, 2$ , and  $Np_1 p_2/m \rightarrow 1$ , as  $N \rightarrow \infty$ . This gives us  $V_1$  as an estimator for  $\text{Var}(\hat{N})$ .

**Example 2.** Suppose that the pairs  $(p_{1i}, p_{2i}), i = 1, \dots, N$ , are independent in the sense that the distribution of  $p_{1i}$ 's is the same for each distinct value of the  $p_{2i}$ 's. Then,  $\bar{p}_{12} = \bar{p}_1 \bar{p}_2, \bar{S}_3 = \bar{S}_1 \bar{S}_2, \bar{S}_4 = \bar{p}_2 \bar{S}_1, \bar{S}_5 = \bar{p}_1 \bar{S}_2$ . Substituting into the Lemma we get

$$\text{Var}(\hat{N}) = N \left( \frac{1}{\bar{p}_1 \bar{p}_2} - \frac{1}{\bar{p}_2} - \frac{1}{\bar{p}_1} - \frac{\bar{S}_1 \bar{S}_2}{\bar{p}_1^2 \bar{p}_2^2} + \frac{\bar{S}_1}{\bar{p}_1^2} + \frac{\bar{S}_2}{\bar{p}_2^2} \right)$$

$$= N \left( \frac{\bar{q}_1 \bar{q}_2}{\bar{p}_1 \bar{p}_2} - cv(p_{1i})^2 cv(p_{2i})^2 \right),$$

where  $cv(p_{ji}) = (\bar{S}_j - \bar{p}_j^2)/\bar{p}_j$ , is the coefficient of variation of the  $p_{ji}$ 's,  $j = 1, 2$ . Obviously,  $\text{Var}(\hat{N}) \leq N\bar{q}_1 \bar{q}_2/(\bar{p}_1 \bar{p}_2)$ . A comparison with Example 1 shows that  $V_1$  is a conservative estimator of  $\text{Var}(\hat{N})$  (i.e.,  $V_1$  is asymptotically too large), when  $p_{1i}$ 's are independent of  $p_{2i}$ 's. Another way of saying this is that, given the means  $\bar{p}_j, j = 1, 2$ , the largest value of the variance is obtained at homogeneity. This is analogous to the variance of the number of successes in Bernoulli trials with variable probabilities of success, cf. Feller 1968, pp. 230-231. A comparison with Example 1 shows that  $V_1$  is a conservative estimator of  $\text{Var}(\hat{N})$  (i.e.,  $V_1$  is asymptotically too large), when the pairs  $(p_{1i}, p_{2i})$  are independent. Note that the independence condition implies that  $C = 0$ .

When the probabilities are not independent, the classical variance estimator is not guaranteed to be conservative. A conservative estimator exists, however. It is obtained by majorizing  $\text{Var}(\hat{N})$  by a quantity that can be estimated in terms of the observable variables. We prove in the Appendix the following general proposition.

**Proposition 1.** A conservative estimator of  $\text{Var}(\hat{N})$  is

$$V_2 = (n_1^2 n_2^2 + n_2^2 m u_1 + n_1^2 m u_2)/m^3,$$

where  $u_j = n_j - m, j = 1, 2$ .

### 3. GAUSSIAN HETEROGENEITY

We shall now turn to a special case in which the sample moments of the pairs  $(p_{1i}, p_{2i})$ ,  $i = 1, \dots, N$ , agree with those of a bivariate normal, or Gaussian, distribution. This will permit a much sharper specification of a conservative variance estimator than the one obtained in the general case above. Assume that

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} v_1^2 \mu_1^2 & \rho v_1 v_2 \mu_1 \mu_2 \\ \rho v_1 v_2 \mu_1 \mu_2 & v_2^2 \mu_2^2 \end{bmatrix}\right),$$

where  $|\rho| < 1$ , and  $0 < \mu_j < 1$ ,  $j = 1, 2$ . Note that  $v_j$ 's can be interpreted as the coefficients of variation of the distributions of  $p_{ji}$ 's. Write  $\bar{S}_j = S_j/N$  for  $j = 1, \dots, 5$ , as before. Then substitute the moments of the bivariate normal distribution into Lemma 1 as follows,

$$\bar{p}_j = E[X_j] = \mu_j, j = 1, 2;$$

$$\bar{S}_j = E[X_j^2] = \mu_j^2(1 + v_j^2), j = 1, 2;$$

$$\bar{p}_{12} = E[X_1 X_2] = \mu_1 \mu_2 (1 + \rho v_1 v_2);$$

$$\bar{S}_3 = E[X_1^2 X_2^2] = \mu_1^2 \mu_2^2 (1 + v_1^2 + v_2^2 + 4\rho v_1 v_2 + (2\rho^2 + 1)v_1^2 v_2^2);$$

$$\bar{S}_4 = E[X_1^2 X_2] = \mu_1^2 \mu_2 (1 + 2\rho v_1 v_2 + v_1^2);$$

$$\bar{S}_5 = E[X_1 X_2^2] = \mu_1 \mu_2^2 (1 + 2\rho v_1 v_2 + v_2^2).$$

Straightforward, but slightly tedious calculations prove then the following proposition (details omitted).

**Proposition 2.** With the above assumptions

$$\text{Var}(\hat{N}) = A_1 A_2 + RN,$$

where

$$A_1 = N/(1 + \rho v_1 v_2)^2;$$

$$A_2 = [1 - (\mu_1 + \mu_2)(1 + \rho v_1 v_2 + \mu_1 \mu_2(1 + \rho v_1 v_2)^2)/[\mu_1 \mu_2(1 + \rho v_1 v_2)^2];$$

$$R = (2\rho v_1 v_2 + 3\rho^2 v_1^2 v_2^2 - \rho^2 v_1^4 v_2^2 - \rho^2 v_1^2 v_2^4 - v_1^2 v_2^2)/(1 + \rho v_1 v_2)^4.$$

We can evaluate the classical variance estimator  $V_1 = n_1 n_2 u_1 u_2 / m^3$  using this result. Note first that  $\{n_1 n_2 / m\} / A_1 \rightarrow 1$ , as  $N \rightarrow \infty$ . Similarly,  $\{u_1 u_2 / m^2\} / A_2 \rightarrow 1$ . This proves the following corollary to Proposition 2:  $(V_1 - \text{Var}(\hat{N})) / N \rightarrow -R$ , as  $N \rightarrow \infty$ . For example, if  $\rho = 0$ , then  $-R = v_1^2 v_2^2$ , so that  $V_1$  is seen to overestimate the asymptotic variance. This is in accordance with Example 2.

How reasonable is the assumption of Gaussian moments? Certainly the capture probabilities cannot have strictly Gaussian distributions, because the Gaussian distribution always puts some probability mass outside the unit interval. On the other hand, suppose we generate the  $p_{ji}$ 's by taking  $\text{logit}(p_{ji}) = a_j + b_j Y_{ji}$ , where the pairs  $(Y_{1i}, Y_{2i})$  are a sample from a bivariate normal distribution with mean zero, unit variances, and correlation  $\rho$ . If we have the relations  $a_j = -\text{logit}(\mu_j)$ ,  $j = 1, 2$ , and  $b_j = v_j(1 + \mu_j)^2$ , then the assumption of Gaussian moments is approximately true. In fact, even the distribution of the pairs  $(p_{1i}, p_{2i})$  is in that case approximately bivariate Gaussian.

Let us consider the adequacy of  $V_1$  further, under the Gaussian moments. The fact that probabilities are constrained to be between 0 and 1 means that  $\mu_j$ 's are between zero and one. Moreover, to be sure that most of the probability mass is in the unit square, let us assume that  $0 < v_j \leq 1/2$ ,  $j = 1, 2$ . If  $\mu_j$ 's are close to one, a much smaller upper bound would be needed. Assume now that  $\rho \leq 0$ . Then, one can show that

$$-R \geq (\rho^2 v_1^4 v_2^2 + \rho^2 v_1^2 v_2^4) / (1 + \rho v_1 v_2)^4 > 0,$$

so that  $V_1$  overestimates  $\text{Var}(\hat{N})$  for  $\rho \leq 0$  also. Note that by continuity  $V_1$  must overestimate  $\text{Var}(\hat{N})$  for some positive values of  $\rho$ , as well.

One can show that  $R = R(\rho)$  is an increasing function of  $\rho$  for at least  $\rho > 0$ . In the limit we have

$$-R(\rho) \rightarrow (-2v_1 v_2 - 2v_1^2 v_2^2 + v_1^4 v_2^2 + v_1^2 v_2^4) / (1 + v_1 v_2)^4,$$

as  $\rho \rightarrow 1$ . When  $0 < v_j \leq 1/2$ ,  $j = 1, 2$ , the smallest value of the above limit occurs at  $v_1 = v_2 = 1/2$ . The minimum value is  $-152/625 > -1/4$ . Consequently, for  $\rho > 0$ ,  $V_1$  can either underestimate or overestimate  $\text{Var}(\hat{N})$ .

The practical implications of the above results are as follows. First, if  $\rho \leq 0$ , then  $\hat{N}$  is either consistent or it overestimates  $N$  and  $V_1$  gives an overestimate of the variance, so we can calculate a conservative upper confidence limit for  $N$ . When  $\rho > 0$ ,  $\hat{N}$  gives an underestimate of  $N$ . If, in addition,  $\rho$  is small, then  $V_1$  gives an overestimate, and we can get a conservative lower confidence limit for  $N$ . Obviously, these are rather special circumstances that one would not expect to be of wide practical utility.

Under the present model the asymptotic bias of  $V_1$  is  $> -N/4$  for all values of  $\rho$ . We can derive a conservative variance estimator by noting that in the Gaussian case the asymptotic relative bias of  $\hat{N}$  is  $-\rho v_1 v_2 / (1 + \rho v_1 v_2) \geq -1/5$ . Hence, asymptotically  $5\hat{N}/4 \geq N$ . A conservative estimator of  $\text{Var}(\hat{N})$  is, for example,  $V_3 = V_1 + 5\hat{N}/16$ . This can be much smaller than  $V_2$  indicating that the Gaussian assumption is a very powerful one.

#### 4. AN APPLICATION TO OCCUPATIONAL DISEASE REGISTRATION DATA

To get an idea of how large the biases may be in practice, let us look at occupational disease data from Finland as an example. The Finnish Register of Occupational Diseases has been in operation since 1964. It is kept by the Institute of Occupational Health in Helsinki. Since 1975 the number of new cases reported to the Register has varied from about 4,000 to over 7,000 annually (0.2 – 0.4 % of the employed population). Noise-induced hearing loss, diseases caused by repetitive or monotonous work (epicondylitis, bursitis, tendinivaginitis), and skin diseases



are the major diagnostic groups (*cf.* Vaaranen *et al.* 1985). The Register can be viewed as a dual registration system, because each case of disease should, under existing regulations, be reported to the Register both from the appropriate insurance company and the examining physician.

It is likely that the probability of reporting a case depends on diagnosis, for example. Indeed, based on data from the year 1981 we get the following statistics. Reports from the insurance companies,  $n_1 = 3,769$ ; reports from the physicians,  $n_2 = 3,053$ ; and cases reported from both sources,  $m = 1,591$ . Thus the usual dual registration estimate is  $\hat{N} = 7,232$  with  $V_1^{1/2} = 97$ ,  $V_2^{1/2} = 222$ , and  $V_3^{1/2} = 108.0$ . The closeness of  $V_3$  to  $V_1$  is striking. Stratifying the data into four categories by diagnosis (the three diagnostic groups mentioned above, and the remaining "other" category) yields the following estimates. Noise-induced hearing loss:  $\hat{N} = 2,230$ ,  $V_1^{1/2} = 33.4$ ,  $V_2^{1/2} = 47.2$ , and  $V_3^{1/2} = 42.6$ ; diseases caused by repetitive or monotonous work:  $\hat{N} = 3,572$ ,  $V_1^{1/2} = 201.4$ ,  $V_2^{1/2} = 303.8$ , and  $V_3^{1/2} = 204.2$ ; skin diseases:  $\hat{N} = 1,441$ ,  $V_1^{1/2} = 30.9$ ,  $V_2^{1/2} = 86.2$ , and  $V_3^{1/2} = 37.5$ ; other diseases  $\hat{N} = 1,015$ ,  $V_1^{1/2} = 32.7$ ,  $V_2^{1/2} = 79.1$ , and  $V_3^{1/2} = 37.2$ . Adding the results yields the following estimates for the total number of diseases:  $\hat{N} = 8,258$ ,  $V_1^{1/2} = 209.0$ ,  $V_2^{1/2} = 340.3$ , and  $V_3^{1/2} = 215.2$ . We see that diseases caused by repetitive or monotonous work are underreported to a particularly great extent.

The analysis was extended further by stratifying the data by diagnosis (4 categories), insurance company (11 categories), and main groups of industry (7 categories). *A priori*, these factors could be thought to have an influence on reporting probabilities. However, the stratification did not alter the point estimate materially. It did increase the estimated standard deviations by over a third, apparently because some of the strata became very small. We conclude that the bias in the point estimator caused by diagnosis is the dominant source of error in the classical estimator in this application.

The same data were further analyzed using a logistic regression technique that allows us to take into account observable population heterogeneity due to both discrete and continuous explanatory variables. In this application age was shown to have an effect on reporting probabilities within the diagnostic groups for one source of information, but not for the other. Therefore, the point estimates remained unchanged and the conclusion regarding the role of diagnosis could not be refuted (Alho 1990).

## 5. DISCUSSION

Our theoretical results indicate that the usual variance estimator  $V_1$  is conservative when the two registration systems are negatively correlated or independent. By continuity the estimator may be conservative also when the correlation is positive but small. Under high positive correlation  $V_1$  gives too low values. We introduced an alternative estimator  $V_2$ , which is conservative under arbitrary population heterogeneity. However, it appears to be unduly conservative in view of the numerical comparisons with  $V_3$ , which is guaranteed to be conservative under Gaussian heterogeneity. The closeness of  $V_3$  to  $V_1$  suggests that, in practice,  $V_1$  may be fairly robust against population heterogeneity.

Unfortunately, even the use of the conservative estimator  $V_2$  would not have been sufficient to cover the bias in the classical point estimator in our empirical example. Perhaps this was to be expected, since the bias of  $\hat{N}$  and the degree of overestimation provided by  $V_2$  are both of order  $N$ . Hence, the use of  $V_2$  inflates the width of a confidence interval by a factor of order  $N^{1/2}$  only. Therefore,  $V_2$  can compensate for the bias of  $\hat{N}$ , if the bias is small, or if  $N$  itself is small. Hence, it seems that the successful application of the dual registration method requires that either we have roughly uncorrelated registration systems, or that the heterogeneity is

observable. In the latter case we may use stratification as suggested already by Sekar and Deming (1949), or logistic regression modeling as suggested by Huggins (1989) and Alho (1990), to adjust for the bias of the classical estimator of population size.

### ACKNOWLEDGEMENT

The author would like to thank Bruce Spencer and an anonymous referee for comments that helped to improve the presentation. Part of the empirical results were first presented at the 11th Nordic Conference on Mathematical Statistics in Uppsala, Sweden, in June 1986.

### APPENDIX

**Proof of Lemma 1.** Apply a linear Taylor-series development to  $\hat{N} = n_1 n_2 / m$  at  $E[n_1] E[n_2] / E[m] = N \bar{p}_1 \bar{p}_2 / \bar{p}_{12}$ , or

$$\hat{N} \approx \frac{N \bar{p}_1 \bar{p}_2}{\bar{p}_{12}} + \frac{\bar{p}_2}{\bar{p}_{12}} (n_1 - N \bar{p}_1) + \frac{\bar{p}_1}{\bar{p}_{12}} (n_2 - N \bar{p}_2) - \frac{\bar{p}_1 \bar{p}_2}{\bar{p}_{12}^2} (m - N \bar{p}_{12}).$$

Hence, we have

$$\begin{aligned} E \left[ \left( \hat{N} - \frac{N \bar{p}_1 \bar{p}_2}{\bar{p}_{12}} \right)^2 \right] &\approx \left( \frac{\bar{p}_2}{\bar{p}_{12}} \right)^2 \text{Var}(n_1) + \left( \frac{\bar{p}_1}{\bar{p}_{12}} \right)^2 \text{Var}(n_2) + \left( \frac{\bar{p}_1 \bar{p}_2}{\bar{p}_{12}^2} \right)^2 \text{Var}(m) \\ &\quad - \frac{\bar{p}_1 \bar{p}_2^2}{\bar{p}_{12}} \text{Cov}(n_1, m) - 2 \frac{\bar{p}_1^2 \bar{p}_2}{\bar{p}_{12}^3} \text{Cov}(n_2, m). \end{aligned}$$

Under our independence assumptions  $\text{Var}(n_j) = N \bar{p}_j - S_j$ ,  $j = 1, 2$ ;  $\text{Var}(m) = N \bar{p}_{12} - S_3$ ,  $\text{Cov}(n_1, m) = -S_4 + N \bar{p}_{12}$ , and  $\text{Cov}(n_2, m) = -S_5 + N \bar{p}_{12}$ . Substituting these into the mean squared error gives the result.

**Proof of Proposition 1.** We ignore the negative term containing  $S_3$  in Lemma 1. Since  $0 < p_{ji} < 1$ , we have  $S_4 < N \bar{p}_{12}$ , and  $S_4 < S_1$ . Therefore,

$$\frac{2 \bar{p}_1 \bar{p}_2^2}{\bar{p}_{12}^3} S_4 < \frac{\bar{p}_1 \bar{p}_2^2}{\bar{p}_{12}^3} N \bar{p}_{12} + \frac{\bar{p}_2^2}{\bar{p}_{12}^2} S_1 + \left( \frac{\bar{p}_1 - \bar{p}_{12}}{\bar{p}_{12}} \right) \frac{\bar{p}_2^2}{\bar{p}_{12}^2} N \bar{p}_{12}.$$

Similarly,

$$\frac{2 \bar{p}_1^2 \bar{p}_2}{\bar{p}_{12}^3} S_5 < \frac{\bar{p}_1^2 \bar{p}_2}{\bar{p}_{12}^3} N \bar{p}_{12} + \frac{\bar{p}_1^2}{\bar{p}_{12}^2} S_2 + \left( \frac{\bar{p}_2 - \bar{p}_{12}}{\bar{p}_{12}} \right) \frac{\bar{p}_1^2}{\bar{p}_{12}^2} N \bar{p}_{12}.$$

Substituting these bounds to the expression of Lemma 1 we get

$$\text{Var}(\hat{N}) < \frac{\bar{p}_1^2 \bar{p}_2^2}{\bar{p}_{12}^3} N + \frac{(\bar{p}_1 - \bar{p}_{12}) \bar{p}_2^2}{\bar{p}_{12}^3} N + \frac{(\bar{p}_2 - \bar{p}_{12}) \bar{p}_1^2}{\bar{p}_{12}^3} N.$$

Estimating  $N \bar{p}_j$  by  $n_j$ ,  $j = 1, 2$ ; and  $N \bar{p}_{12}$  by  $m$  we get the result.

## REFERENCES

- ALHO, J. (1990). Logistic regression in capture-recapture models. *Biometrics*, 46, 623-635.
- BURNHAM, P.K., and OVERTON, W.S. (1979). Robust estimation of population size when capture probabilities vary among animals. *Ecology*, 60, 927-936.
- FELLER, W. (1968). *An Introduction to Probability Theory and Its Applications*, (Vol. I, 3<sup>rd</sup> ed.). New York: Wiley.
- HUGGINS, R.M. (1989). On the statistical analysis of capture experiments. *Biometrika*, 76, 133-140.
- SEBER, G.A.F. (1982). *The Estimation of Animal Abundance*, (2<sup>nd</sup> ed.). New York: Griffin.
- SEKAR, C.C., and DEMING, W.E. (1949). On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association*, 44, 101-115.
- VAARANEN, V., VASAMA, M., and ALHO, J. (1985). Occupational diseases in Finland in 1984. Reviews 11, Institute of Occupational Health, Helsinki.



## Combining Estimates from Surveys

PODURI S.R.S. RAO and I.M. SHIMIZU<sup>1</sup>

### ABSTRACT

For estimating the proportion and total of an item for the present occasion, independent estimates at the current and previous occasions are combined through three different procedures. In the first one, trend over the occasions is utilized. For the second one, the One-Way Random Effects Model is employed. The third procedure uses the Empirical Bayes approach. All the three procedures are seen to perform better than the sample estimates obtained from the data of the current occasion alone. Advantages of these methods and their limitations are discussed. All the procedures are illustrated with the data from the National Health Discharge Survey.

**KEY WORDS:** Trend; Weighted least squares; Random effects; Improved estimation; Biases; Mean square errors.

### 1. INTRODUCTION

In several national surveys, independent samples are obtained at successive time periods. In this paper, information from the past surveys is utilized to improve the estimates for the current period. For the sake of illustration, we consider the National Health Discharge Survey (NHDS) in the U.S. In this survey, which has been recently redesigned, a three stage sampling design is used with geographical regions as the Primary Sampling Units (PSU's) at the first stage. Hospitals and discharges are selected at the second and third stages respectively. The survey collects information on various items of the patients like age, sex, racial characteristics, length of stay, diagnosis, and surgical and non-surgical procedures. The selected PSU's and hospitals remain in the study for a certain number of years. Independent samples of discharges are obtained every year from the selected hospitals. Shimizu (1987) presents further details of the redesign of the NHDS.

At present, for a given hospital, estimates of the proportions for the different items for the current year are obtained only from the data of this year. National estimates are obtained by suitably weighting these proportions with the reciprocals of the probabilities of selection of the hospitals and the PSU's. However, Bean (1987) found that for most of the items the estimates are somewhat correlated over the years. For the sake of illustration, sample proportions obtained from the NHDS for 1977-86 for Acute Myocardial Infraction (AMI) and Mental Disorders (MDS) are presented in Table 1 for three hospitals and they are exhibited in Figures 1 and 2. Examination of the proportions for these three and 17 more hospitals suggested that the inclusion of past information can increase the precision of the estimates for the current year.

It should be cautioned that the sample proportions in Table 1 or Figures 1 and 2 should not be used to make inferences regarding the increase or decrease of AMI or MDS in the entire population.

<sup>1</sup> Poduri S.R.S Rao, Department of Statistics, Hylan 703, University of Rochester, Rochester NY, 14618 U.S.A., and I.M. Shimizu, National Center for Health Statistics, Office of Research and Methodology 1-68, 3700 East-West Highway, Hyattsville MD, 20782, U.S.A.

**Table 1**  
Data from the National Health Discharge Survey for 1977-86  
Sample totals and proportions for Acute Myocardial  
Infraction (AMI) and Mental Disorders  
(MDS) for three hospitals

Year	No. of discharges <i>N</i>	Sampled No. of discharges <i>n</i>	AMI		MDS	
			Total	Sample proportion	Total	Sample proportion
1977	9,416	276	5	.018	37	.13
1978	10,234	266	7	.026	24	.09
1979	9,354	294	9	.031	39	.13
1980	10,372	327	9	.028	41	.13
1981	10,712	342	8	.023	45	.13
1982	10,683	309	9	.029	43	.14
1983	10,935	360	7	.019	46	.15
1984	10,090	330	6	.018	50	.15
1985	10,431	297	8	.027	41	.14
1986	10,247	264	4	.015	35	.13
1977	6,720	474	9	.019	18	.04
1978	6,710	470	14	.030	25	.05
1979	6,970	495	8	.016	28	.06
1980	6,794	466	14	.030	29	.06
1981	7,055	486	9	.019	34	.07
1982	6,265	442	9	.020	24	.05
1983	6,234	442	10	.023	28	.06
1984	6,221	439	9	.021	15	.03
1985	6,063	375	8	.021	19	.05
1986	5,781	371	4	.011	12	.03
1977	6,400	606	21	.0347	41	.0677
1978	6,286	635	23	.0362	42	.0661
1979	6,494	554	12	.0217	27	.0487
1980	6,813	571	17	.0298	25	.0438
1981	7,430	729	14	.0192	32	.0439
1982	7,267	712	20	.0281	39	.0548
1983	7,110	694	23	.0331	43	.0620
1984	7,268	718	35	.0487	29	.0404
1985	6,716	657	19	.0289	45	.0685
1986	6,464	655	21	.0321	33	.0504

In this article, we examine three procedures for improving the estimates for a specified hospital by utilizing the information from the current and the previous years. In the first method, estimates of the proportions are obtained from the linear trend over the years and the Weighted Least Squares Method. If there is a significant positive or negative trend over the years, this method will have higher precision than the sample estimate of the current period. If the trend is not pronounced, the increase in precision will be negligible, as expected.

For the second procedure, the One-Way Random Effects Model with unequal variances is used to combine the information. Yates and Cochran (1938) and Cochran (1954) suggested this type of procedure for combining information from experiments conducted at different time periods and locations. While the Analysis of Variance (ANOVA) method had been used

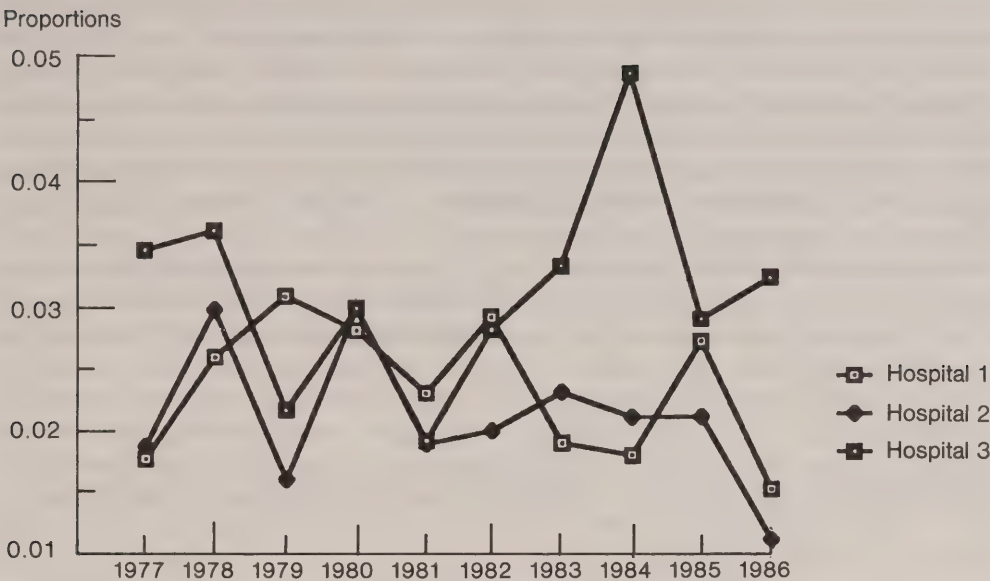


Figure 1. Proportions for AMI: 1977-86

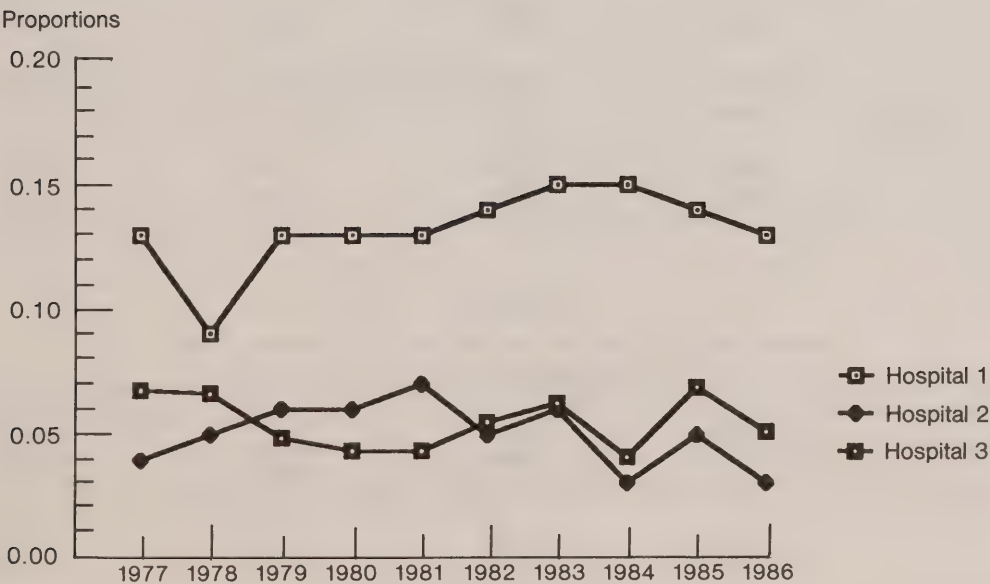


Figure 2. Proportions for MDS: 1977-86



for quite some time for this purpose, C.R. Rao (1970) suggested the Minimum Norm Quadratic Unbiased Estimation (MINQUE) and demonstrated its advantages. P.S.R.S. Rao, Kaplan and Cochran (1981) examined the relative merits of the ANOVA, MINQUE and several related procedures. We have employed the estimation procedures related to these methods. The estimate for the proportion obtained by any of these procedures is a weighted combination of the estimates of the different time periods. The weights depend on both the between and within variances of the time periods. In the third procedure, the Empirical Bayes approach is used to estimate the proportions for the current period.

We denote the above three procedures by TR, VC and EB respectively. The notation is presented in Section 2. The sample estimator for the proportion and its variance are given in Section 3. The above three estimation procedures along with the expressions for their Standard Errors (S.E.'s) are presented in Sections 4, 5 and 6. We have used these expressions to compute for 1986 the sample proportions, the above three types of estimates, and their S.E.'s for 20 hospitals in the NHDS. These estimates for the three hospitals mentioned earlier are presented in Table 2 for AMI and Table 3 for MDS. Results from the entire study are described in Section 7. The final section contains a discussion of the results and topics for further research.

**Table 2**  
Estimates of the Proportions for 1986 and S.E.'s (bottom figures)  
for Acute Myocardial Infractions (AMI)

Hospital	Sample proportion	Trend estimate	Variance components estimate	Bayes estimate
1	.0152	.0196	.0224	.0224
	.0070	.0046	.0026	.0003
2	.0108	.0162	.0204	.0203
	.0048	.0036	.0031	.0003
3	.0321	.0319	.0304	.0309
	.0060	.0038	.0028	.0037

**Table 3**  
Estimates of the Proportions for 1986 and S.E.'s (bottom figures)  
for Mental Disorders (MDS)

Hospital	Sample proportion	Trend estimate	Variance components estimate	Bayes estimate
1	.1326	.1431	.1292	.1292
	.0205	.0115	.0060	.0010
2	.0323	.0437	.0500	.0427
	.0087	.0056	.0039	.0057
3	.0504	.0496	.0534	.0523
	.0080	.0049	.0032	.0048

It should be mentioned that for the problem considered in this paper, the samples are drawn independently at the different time periods. Secondly, the population proportions for the previous periods are not known. Because of these reasons, the usual ratio and regression methods cannot be employed to improve the accuracy of the estimators for the current period. For the same reasons, the estimation procedures suggested in the literature for the rotation sampling schemes cannot be used in this situation. In spite of these difficulties, the three methods considered in this paper can be used to estimate the population quantities with a high accuracy. When summary figures at the different periods are available, public and private users can obtain these estimates and their standard errors without much difficulty. These procedures can also be used when there is nonresponse during some years – some of the hospitals do not provide information to the survey during some years.

2. NOTATION

We present in this section the notation for a selected PSU. Let  $y_{itj}$  denote the  $j$ th observation on the sampled discharge on an item like the number of surgical cases at time  $t = (1, 2, \dots, T)$ , from the  $i$ th hospital,  $i = (1, 2, \dots, K)$ , which has  $N_{it}$  discharges. Note that  $K$  may change over the years due to nonresponse or the addition of new hospitals.

The total and mean at time  $t$  are

$$Y_{it} = \sum_1^{N_{it}} y_{itj} \tag{1}$$

and

$$\bar{Y}_{it} = Y_{it}/N_{it}. \tag{2}$$

The total and mean of the sample of size  $n_{it}$  from the  $N_{it}$  discharges are

$$y_{it} = \sum_1^{n_{it}} y_{itj} \tag{3}$$

and

$$\bar{y}_{it} = y_{it}/n_{it}. \tag{4}$$

To estimate the total number and proportion for a specified item, let  $y_{itj} = 1$  if the observation belongs to that item, and zero otherwise. With this notation, the total and proportion for an item at time  $t$  can be written as  $A_{it}$  and  $P_{it} = A_{it}/N_{it}$ . Note that  $P_{it}$  is the same as  $\bar{Y}_{it}$ .

In the following four sections, for the sake of convenience, we suppress the subscript  $i$  and describe the estimators for a given hospital.

3. SAMPLE PROPORTION

An unbiased estimator of the proportion  $P_t$  for an item like AMI or MDS is

$$\hat{P}_t = a_t/n_t, \tag{5}$$

where  $a_t$  is the number of cases of that item observed in the  $n_t$  sample discharges. The variance of  $\hat{P}_t$  and its unbiased estimator are

$$V(\hat{P}_t) = \frac{N_t - n_t}{N_t - 1} \frac{P_t(1 - P_t)}{n_t} \quad (6)$$

and

$$v(\hat{P}_t) = (1 - f_t) \frac{\hat{P}_t(1 - \hat{P}_t)}{n_t - 1}, \quad (7)$$

where  $f_t = n_t/N_t$ . Note that  $\hat{P}_t$  is the same as  $\bar{y}_t = \sum_1^{n_t} y_{ij}/n_t$ .

#### 4. LINEAR TREND

The sample observations  $y_{ij}$ ,  $j = (1, 2, \dots, n_{ij})$  can be written as

$$y_{ij} = \mu_t + \epsilon_{ij}, \quad (8)$$

where  $\mu_t$  is the mean for the  $i$ th hospital at the  $t$ th period, and  $\epsilon_{ij}$  is the random error with expectation zero and variance  $\sigma_t^2 = P_t(1 - P_t)$ . Since the samples are drawn independently during each year, the errors  $\epsilon_{ij}$  are uncorrelated from one year to another.

With the assumption of a linear trend, the sample mean can be expressed as

$$\bar{y}_t = \alpha + \beta x_t + \bar{\epsilon}_t, \quad (9)$$

where  $x_t = t$  and  $\bar{\epsilon}_t = \sum_1^{n_t} \epsilon_{ij}/n_t$ . Further,  $V(\bar{\epsilon}_t) = (N_t - n_t)\sigma_t^2/(N_t - 1)n_t = 1/W_t$ . Note that with the zero-one notation,  $\bar{y}_t$  is the same as  $\hat{P}_t$ . The WLS estimators of  $\beta$  and  $\alpha$  are

$$\hat{\beta} = \frac{\sum W_t(x_t - \bar{x})\bar{y}_t}{\sum W_t(x_t - \bar{x})^2} \quad (10)$$

and

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}, \quad (11)$$

where  $\bar{x} = \sum W_t x_t / \sum W_t$  and  $\bar{y} = \sum W_t \bar{y}_t / \sum W_t$ .

Estimator of  $\mu_t$  is

$$\begin{aligned} \hat{\mu}_t &= \hat{\alpha} + \hat{\beta}x_t \\ &= \bar{y} + \hat{\beta}(x_t - \bar{x}). \end{aligned} \quad (12)$$

This is the Trend Estimator (TR). Estimators of this type for infinite populations have been examined in the literature; see Carroll and Rupert (1988), for instance.



We have obtained the estimate of  $\mu_t$  from this expression by replacing  $W_t$  with  $w_t = (1 - f_t) \hat{\sigma}_t^2/n_t$ , where  $\hat{\sigma}_t^2 = \hat{P}_t(1 - \hat{P}_t)$ . If it can be assumed that for large  $N_t$  the distribution of  $y_{tj}$  is normal,  $\bar{y}_t$  will be independent of  $w_t$ . In this case, the expression in (12) remains unbiased for  $\mu_t$ . Even if the assumption of normality is not valid, it can be seen that  $w_t$  approaches  $W_t$  for large  $n_t$  and hence the expression in (12) with the estimated weights approaches  $\mu_t$ .

The variance of the above estimator is

$$V(\hat{\mu}_t) = \frac{1}{\sum W_t} + \frac{(x_t - \bar{x})^2}{\sum W_t(x_t - \bar{x})^2}. \tag{13}$$

We have estimated this variance by replacing  $W_t$  by  $w_t$ . The bias in the resulting estimator will be small for large  $n_t$ .

For the illustration in this article,  $t = (1, 2, \dots, 10)$ , that is,  $T = 10$ . For 1986, we have found the estimate for the proportion of an item and its S.E. from (12) and (13) with  $x_t = 10$ .

5. VARIANCE COMPONENTS MODEL

Examination of the proportions for the AMI and MDS of the 20 hospitals for the ten years showed no specific linear or nonlinear trend. For all of them the patterns somewhat resembled those of the three hospitals, presented in Figures 1 and 2. These observations indicated that the proportion for AMI or MDS for the current year can be obtained by combining the information from all the ten years. The One-Way Random Effects Model can be used for this purpose.

The model in (8) can be written as

$$\begin{aligned} y_{tj} &= \mu + (\mu_t - \mu) + \epsilon_{tj} \\ &= \mu + \alpha_t + \epsilon_{tj}. \end{aligned} \tag{14}$$

If  $\mu_t$  is considered to be randomly drawn from a population with mean  $\mu$ , the random effect  $\alpha_t$  will have mean zero and variance  $\sigma_\alpha^2$ . It is assumed to be independent of  $\epsilon_{tj}$ . The sample mean (proportion) can now be written as

$$\bar{y}_t = \mu + \alpha_t + \bar{\epsilon}_t, \tag{15}$$

where  $\bar{\epsilon}_t$  has mean zero and variance  $(1 - f_t) \sigma_\alpha^2/n_t$ . Thus, from (15),

$$V(\bar{y}_t) = \sigma_\alpha^2 + (N_t - n_t)\sigma_t^2/(N_t - 1)n_t = \frac{1}{U_t}. \tag{16}$$

The WLS estimator of  $\mu$  is

$$\hat{\mu} = \frac{\sum U_t \bar{y}_t}{\sum U_t}. \tag{17}$$

This is the Variance Components Estimator (VC) and its variance is

$$V(\hat{\mu}) = 1/\sum U_i. \quad (18)$$

For obtaining the mean in (17) and its variance in (18), we have replaced  $\sigma_t^2$  by its estimate  $\hat{P}_t(1 - \hat{P}_t)$ . Procedures like the ANOVA and MINQUE are available for estimating  $\sigma_\alpha^2$ . The MINQUE depends on the *a priori* values  $r_t$  of  $(\sigma_t^2/\sigma_\alpha^2)$ . A related procedure called the Unweighted Sums of Squares (USS) method does not depend on  $r_t$  and it is described below. P.S.R.S. Rao, Kaplan and Cochran (1981) found that this method provides estimates for  $\sigma_\alpha^2$  comparable to the ANOVA and MINQUE, unless  $n_t$  or  $r_t$  is very small. The USS is computationally less cumbersome than the MINQUE. With  $\bar{y}^* = (\sum \bar{y}_t)/T$ , from (15),

$$E[\sum (\bar{y}_t - \bar{y}^*)^2] = (T - 1)\sigma_\alpha^2 + (T - 1)(\sum v_t)/T, \quad (19)$$

where  $v_t = (N_t - n_t)P_t(1 - P_t)/(N_t - 1)n_t$ . The USS estimator for  $\sigma_\alpha^2$  is

$$\hat{\sigma}_\alpha^2 = \sum (\bar{y}_t - \bar{y}^*)^2/(T - 1) - (\sum \hat{v}_t)/T, \quad (20)$$

where  $\hat{v}_t = (1 - f_t)\hat{P}_t(1 - \hat{P}_t)/(n_t - 1)$ . If  $N_t$  is large relative to  $n_t$ , the sampling fraction  $f_t$  can be set to zero. We have estimated  $U_t$  from (16) by estimating  $\sigma_\alpha^2$  from (20) and the second term by  $\hat{v}_t$ . Utilizing this estimate of  $U_t$ , we have estimated  $\mu$  from (17) and its variance from (18). If  $\sigma_\alpha^2$  is much larger than  $v_t$ , the estimator  $\hat{\mu}$  in (17) will be close to  $\bar{y}^*$ . In this case, estimation of  $U_t$  as described above can be expected to have almost no effect on  $\hat{\mu}$ . Since  $\hat{\mu}$  depends only on the relative values of  $U_t$ , this conclusion can be expected to be valid even when  $\sigma_\alpha^2$  is not considerably larger than  $v_t$ . Thus, estimation of  $U_t$  can be expected to result in only a negligible bias for  $\hat{\mu}$ .

As is well-known, all the procedures for estimating  $\sigma_\alpha^2$  unbiasedly can result in negative estimates. In such a case, we have employed the usual practice of substituting a small positive quantity for the negative estimate. In Rao *et al.* (1981) it was found that unless  $\sigma_\alpha^2$  is very small, this adjustment results in only a negligible bias for  $\hat{\sigma}_\alpha^2$  and an insignificant increase in its standard error. Further, unless  $\sigma_\alpha^2$  is small, the difference in the MSE of  $\hat{\mu}$  for the USS and other methods of estimating  $U_t$  was found to be negligible.

## 6. BAYES' ESTIMATOR

The discussion in the beginning of Section (5) suggests that  $\mu_t$  can be assumed to have a prior distribution with mean  $\mu$  and variance  $\sigma_\alpha^2$ . With the assumptions that for large  $N_t$  the distribution of  $y_{ij}$  is normal with mean  $\mu_t$  and variance  $\sigma_t^2$ , and that the prior distribution of  $\mu_t$  is also normal, the Bayes' Estimator for  $\mu_t$  is

$$B_t = E(m_t | \bar{y}_t) = (1 - a_t)\bar{y}_t + a_t\mu, \quad (21)$$

where  $a_t = v_t/(\sigma_\alpha^2 + v_t)$ . The expression for  $v_t$  is the same as given in the previous section.

For given  $\bar{y}_t$ , the variance of the above estimator is

$$V(B_t) = \frac{1}{(1/\sigma_\alpha^2) + (1/v_t)}. \quad (22)$$

With estimates  $\hat{\sigma}_\alpha^2$ ,  $\hat{\sigma}_t^2$  and  $\hat{\mu}$ , the expression in (21) can be written as

$$\hat{B}_t = (1 - \hat{a}_t) \bar{y}_t + \hat{a}_t \hat{\mu}, \quad (23)$$

where  $\hat{a}_t = \hat{v}_t / (\hat{\sigma}_\alpha^2 + \hat{v}_t)$ . This estimator may be called the Empirical Bayes' estimator (EB). Note that  $\hat{\mu}$  is obtained from (17) with  $\hat{\sigma}_\alpha^2$  and  $\hat{v}_t$ . The variance of this estimator may be obtained from (22) by replacing  $\sigma_\alpha^2$  and  $v_t$  with their estimates. For obtaining the EB and its variance, we have estimated  $\sigma_\alpha^2$  and  $v_t$  from the USS procedure described in the previous section.

## 7. PERFORMANCE OF THE ESTIMATORS

We have computed the estimates of  $P_t$  for 1986 for the 20 hospitals through the different procedures described in the previous sections. Since the population values of  $P_t$  are not known, as described earlier, we have found the S.E.'s for the different procedures by substituting the sample proportion  $\hat{P}_t$  in the place of  $P_t$ . Since the sample sizes  $n_t$  are not small, the resulting biases in estimating the variances or S.E.'s of the estimators can be expected to be small.

For the three hospitals, the estimates of  $P_t$  and the S.E.'s of the different procedures are presented in Tables 2 and 3 for AMI and MDS respectively.

As can be seen from these tables, S.E.'s of TR, VC and EB are smaller than the S.E. of the sample proportion. As expected, utilizing the data from the previous periods has helped reduce the S.E. of the estimate for the current period.

Both VC and EB have smaller S.E.'s than TR. However, TR does not require the estimation of  $\sigma_\alpha^2$ . We have found the S.E. of TR to be usually less than 50 percent of the sample proportion.

The EB has smaller S.E. than VC, as expected. Note that VC estimates the overall proportion, whereas EB estimates the proportion of the conditional distribution. The S.E. of the EB becomes close to that of the sample proportion if the sample size is large.

It is interesting to observe from Tables 2 and 3 that for both AMI and MDS the difference between the VC and EB estimates is negligible. The reason for this result is that  $\hat{a}_t$  is close to unity, which indicates that  $\sigma_\alpha^2$  is small relative to  $v_t$ .

The estimates for the total number of cases for 1986 and their S.E.'s can be obtained by multiplying the estimates of the proportions in Tables 2 and 3 by the corresponding number of discharges  $N_t$  given in Table 1.

## 8. DISCUSSION

As described in the above section, the results of this investigation recommend the TR, VC or EB methods for estimating the proportions and totals for the current period.

For estimating the S.E.'s of the different procedures, we have utilized the sample proportions. Further investigation is needed to examine the biases and MSE's of these S.E.'s.



For estimating  $\sigma_\alpha^2$  and  $v_i$ , we have employed the USS. The effects of the ANOVA and the MINQUE procedures for this purpose can also be examined. However, the investigation in Rao *et al.* (1981) showed that different procedures of estimating  $\sigma_\alpha^2$  may not have a significant effect on the estimation of  $\mu$  or its S.E.

Further investigation is needed to determine the effect of the different procedures of estimating the variances on the EB for  $\mu_i$ .

We have substituted a small positive quantity for a negative estimate of  $\sigma_\alpha^2$ . As can be seen, this adjustment may result in a small S.E. for both the VC and EB, and may present too optimistic a view about the estimates of  $\mu$  and  $\mu_i$ . Further examination of this problem is needed.

We have assumed a linear model for the proportion. The logit or probit transformation can be used before using this model. However, large population and sample sizes are needed to justify the estimates that can be obtained through these transformations. The estimates proposed in this article can be obtained by the public and private users by using any simple computer program.

Improved estimates for each hospital are considered in this paper. The national estimates for a given item like AMI or MDS can be obtained by suitably weighting the above estimates by the reciprocals of the probabilities with which the hospitals were selected. Such a procedure is expected to improve the precision of the national estimates.

Time series methods like the ARIMA can be used as suggested for instance by Blight and Scott (1973) and Scott and Smith (1977) for estimating the proportions and total numbers. These methods will result in different models for different items. Secondly, the available package programs for these approaches assume large population sizes and equal error variances, and the same sample sizes for all the time periods. Such assumptions are not satisfied for the problem we have considered in this article. As mentioned in Section 1, the TR, VC and EB methods can also be used when there is nonresponse during some years.

### ACKNOWLEDGEMENTS

The authors would like to thank the Associate Editor and a referee for their helpful comments.

### REFERENCES

- BEAN, J.A. (1987). NHDS variance and covariance estimation of year to year differences. National Center for Health Statistics, research report.
- BLIGHT, B.J.N., and SCOTT, A.J. (1973). A stochastic model for repeated surveys. *Journal of the Royal Statistical Society, Series B*, 35, 61-68.
- CARROLL, R.J., and RUPERT, D. (1988). *Transformation and Weighting in Regression*. New York: Chapman and Hall.
- COCHRAN, W.G. (1954). The combination of estimates from different experiments. *Biometrics*, 10, 101-129.
- RAO, C.R. (1972). Estimation of variance and covariance components in linear models. *Journal of the American Statistical Association*, 67, 112-115.
- RAO, P.S.R.S., KAPLAN, J., and COCHRAN, W.G. (1981). Estimators for the one-way random effects model with unequal error variances. *Journal of the American Statistical Association*, 76, 89-96.

- SCOTT, A.J., and SMITH, T.M.F. (1977). The application of time series methods to the analysis of repeated surveys. *International Statistical Review*, 45, 13-28.
- SHIMIZU, I.M. (1987). Specifications for the redesigned NHDS sample. National Center for Health Statistics, research report.
- YATES, F., and COCHRAN, W.G. (1938). The analysis of groups of experiments. *Journal of Agricultural Sciences*, 28, 556-580.





## RDD Panel Attrition in Two Local Area Surveys

PAUL J. LAVRAKAS, RICHARD A. SETTERSTEN, Jr.  
and RICHARD A. MAIER, Jr.<sup>1</sup>

### ABSTRACT

This paper compares the magnitude and nature of attrition in two separate RDD panel surveys conducted in the City of Chicago (*i.e.* the surveys were independent studies and were not conducted as part of a planned experiment), each with a between-wave lag of approximately one year. For each survey, sampling at Wave 1 was performed via one-stage (*i.e.* simple) random-digit dialing. In Study 1, respondents' names were *not* elicited; thus, when telephone calls were made at Wave 2 of Study 1 interviewers could not ask for respondents by name. Instead, interviewers asked for respondents by using a gender-age identifier. In Study 2, respondent name identifiers were gathered during Wave 1 and were used in Wave 2 re-contact attempts. The magnitude of the attrition in Study 1 (*i.e.* the proportion of Wave 1 respondents not re-interviewed at Wave 2) was 47%, whereas in Study 2 it was 43%: a marginal difference in attrition rates. In both surveys, age, race, education and income were significantly related to attrition. Discussion is presented on the trade-off between minimizing attrition *vs.* minimizing respondent reactivity as potential sources of total survey error. Suggestions for decreasing the size of attrition in RDD panel surveys are discussed.

KEY WORDS: Panel attrition; Random-digit dialing; Telephone surveys.

### 1. INTRODUCTION AND LITERATURE REVIEW

For the past several decades the problem of panel attrition has received only passing attention in the survey methods literature. Published articles either have addressed techniques which can be employed to minimize the size of the attrition from panel studies (*e.g.* Droege and Crambert 1965; Crider, Willets and Bealer 1971; McAllister, Goe and Bulter 1973; Freedman, Thornton and Camburn 1980; and Burgess 1989) or have addressed statistical techniques that may be used to *adjust* for the effects of panel attrition (*e.g.* Lehnen and Koch 1974; Hausman and Wise 1979; Winer 1983; and Lepkowski 1989).

Few articles have reported on the magnitude and nature of the resulting attrition. And, even fewer have dealt with *random* samples of the public which would allow other researchers to estimate what to expect in future general population surveys. An exception was Sobol's (1959) reporting on the attrition that occurred in a five-wave panel studying economic attitude change. At Wave 1, in 1954, a probability sample of the non-institutionalized urban population of the United States was interviewed ( $n = 1,150$ ). Subsequent waves were conducted six, 12, 18, and 33 months later. Compared to the original sample, attrition for each subsequent wave was 17%, 26%, 29% and 39%, respectively.

Sobol reported that, in general, "because of canceling variations, the demographic structure . . . after five rounds of interviewing, remained very similar to that of the original [sample]" (p. 52). Yet there were some significant variations, with a disproportionate number of renters, lower income households, residents of large metropolitan areas, younger (under 25 years) and older (over 64 years) adults, and those not interested in the survey subject matter lost to the panel. Winer (1983) reported results of unpublished studies which generally confirmed Sobol's findings.

<sup>1</sup> Paul J. Lavrakas, Richard A. Settersten, Jr. and Richard A. Maier, Jr., Northwestern University Survey Laboratory, 625 Haven St., Evanston IL 60208 - 4150 USA.

It is important to note that in each of these studies interviewers knew Wave 1 respondents' full names. In fact, in their article on techniques to minimize panel attrition, McAllister *et al.* (1973) stressed the importance of gathering detailed information about the respondent's future whereabouts at the end of the interview, including "complete names and addresses of friends and/or relatives . . . of the respondent" (p. 416).

Although it can be argued that panel attrition is a serious enough problem to prompt researchers to obtain the full name and other identifying information of each Wave 1 respondent, this approach may cause problems of its own. In those instances where a respondent's name is elicited as part of the Wave 1 interview, an explanation is sometimes given that the name is important because the respondent may/will be called back after some specified time to determine if any changes occurred. This raises concerns about "evaluation apprehension" (*i.e.* reactivity) on the part of respondents (*cf.* Crano and Brewer 1973). Whereas some authors explicitly address the trade-off between attrition and reactivity (*e.g.* Sobol 1959), it is implicit in most other articles, that authors typically regard reactivity as less a problem than attrition.

All of the aforementioned research was conducted with personal interviews. But what of panel attrition when telephone surveying is done, including those studies in which Wave 1 respondents' names are not recorded? In particular, what can be expected by a researcher who plans *a priori* to conduct a panel telephone survey and thus ask respondents for name identifiers *vs.* a researcher who does not gather respondent name identifiers, either because he/she explicitly chooses not to or because a decision is made *post hoc* to convert a cross-sectional telephone survey to a panel after Wave 1 interviewing is complete?

In an attempt provide a preliminary perspective on these issues, the present paper reports findings on the magnitude and the nature of attrition in two RDD (two-wave) panel studies conducted in the City of Chicago, each with a between-wave lag of approximately one year. It should be noted that these two surveys were conducted independently of each other, not as part a planned test of RDD attrition. As such, there are various differences in the substantive focus and specific execution of the two surveys, beyond the fact that in Study 2 a name identifier was known for most respondents whereas in Study 1 it was not. We explicitly acknowledge that these differences in focus and execution somewhat limit the conclusions that can be drawn from the comparison of the two studies.

For both surveys, one stage (*i.e.* simple) random-digit dialing was used to sample Wave 1 respondents. In Study 1, respondent names were not asked as part of the Wave 1 interview and respondents were not told that they would be re-contacted. In Study 2, name identifiers were gathered at the completion of the Wave 1 interviews and were used to reach respondents at Wave 2. Respondents most often did not provide their full names, instead giving their first name only or other name identifier, *e.g.* nickname or initials. (Interviewers did not probe for full names so as to not contribute to possible feelings of paranoia on the part of reluctant respondents.)

When respondents' names are not known, how does one go about re-contacting the original respondent? This was a problem faced in 1979 by the first author when trying to determine the efficacy of creating a panel from a 1977 cross-sectional survey. As nothing was found in the published literature to provide guidance, a pilot-test was conducted with a resulting 50 percent of the 1977 respondents re-interviewed by asking for them by *gender and age*.

The results of this pilot-test were encouraging enough to recommend the procedure for use in the first study reported here. In Study 1, interviewers dialed the same telephone numbers as the Wave 1 completions, verified each number whenever the call was answered, and informed the listener that approximately one year ago a person at the telephone number had completed an interview. The original respondent was identified by *gender* (*e.g.* "a man" or "a woman") and by *age* (*e.g.* "in his early twenties" or "in her late sixties").



In Study 2, a name identifier was known for more than eight out of 10 of the Wave 1 respondents. For these respondents, Wave 2 interviewers asked for the respondent using the name identifier, after first verifying the telephone number. For the respondents with no name identifier, interviewers asked for the respondent by using demographic identifiers, as in Study 1.

In reporting the results from these studies, it is our modest intention to shed preliminary light on the magnitude and nature of attrition in RDD panels. Although the results should not be generalized to a national RDD sample, they are suggestive. Given the prevalence of RDD sampling, we believe it is important to build a knowledge-base about the attrition that can be expected in panel studies where Wave 1 sampling is done via random-digit dialing, especially when researchers have no Wave 1 name identifier for respondents. By doing this, we can better consider strategies to reduce the size and effects of this attrition.

## 2. STUDY 1

### 2.1 Methodology

In February, 1983, a city-wide (one-stage) RDD survey was conducted by the Northwestern University Survey Laboratory to gather baseline data for professors who were evaluating a series of community crime prevention programs in Chicago neighborhoods. (The questionnaire took an average of 20 minutes to administer.) Approximately 2,800 telephone numbers were dialed in the process of completing 814 interviews. For each residence contacted, one head-of-household (male or female) was systematically selected as the designated respondent (*cf.* Lavrakas 1987; pp. 99-100). Whenever necessary, Spanish-language questionnaires were administered by bilingual interviewers. Up to seven call-backs were made to hard-to-reach respondents. Of all telephone numbers dialed, 1,247 were found to ring in eligible households (defined by the survey sponsors as English-speaking or Spanish-speaking households with at least one adult 19 years of age or older); those eligibles not interviewed either were unavailable at the time calls were made or refused to participate.

One year later, in February, 1984, the Wave 1 telephone numbers were re-dialed to gather "post-test" data for the evaluation project. In those instances where the telephone was answered within eight call-attempts (across different days and times), the following introduction was read by interviewers:

Hello, is this \_\_\_\_\_ ? My name is \_\_\_\_\_, and I'm calling from Northwestern University. About a year ago (February 1983) we conducted an interview with a \_\_\_\_\_ at this number. May I please speak with (her/him)?

The interviewer first verified the telephone number and then gave her/his own name. The third *blank* contained pre-recorded Wave 1 demographic information (gender and age) about each respondent: *e.g.* "woman in her mid 30s," or "man in his early 70s." For those few respondents who had not given their year of birth at Wave 1, the third blank simply contained the gender identifier, "woman" or "man."

Once the interviewer was speaking to the original respondent he/she continued with the following explanation, before beginning the interview:

The information you gave us last year was a big help in understanding the concerns of residents like yourself. We are calling back now to find out some things about the quality of life in Chicago neighborhoods during the past year.



The purpose of this statement was to reinforce the respondent’s willingness to cooperate with the Wave 2 interviewer by reminding the respondent of his/her cooperation in the Wave 1 survey.

Coinciding with the purpose of this evaluation project, respondents who had moved or changed their telephone numbers were not interviewed at Wave 2. This was due to the need to interview only those persons who resided at the same address as the previous year, since many of questions dealt with perceived neighborhood change since February, 1983.

2.2 Results

Due to a clerical error in processing the Wave 1 questionnaires and call-records, duplicate or incorrect respondent I.D. numbers were assigned to 17 Wave 1 respondents by the survey sponsors’ staff. For the purposes of this paper, these respondents were dropped from our analyses because we could not match correctly their Wave 2 dispositions with their respective Wave 1 data. Thus the following analyses are based on the 797 respondents whose Wave 1/ Wave 2 match was certain.

**The magnitude of the attrition.** As shown in Table 1, approximately one-half of the Wave 1 sample was re-interviewed (53%). Of the 375 respondents who were “lost” to the panel, the greatest proportion was due to telephone numbers that rang in a new household or in an original household from which the respondent had moved; this accounted for approximately 40% of the attrition. Second most frequent were those persons whose Wave 1 telephone number was no longer in service; this accounted for a fourth of those lost. Next in frequency of those lost were respondents who refused in some way. The fourth most prevalent reason for losing respondents were those who were never home during the Wave 2 field period when their telephone was answered, even after eight call-backs; (these 33 persons were verified to be the original respondent by someone else in their household).

**Table 1**  
Disposition of Wave 1 Samples for Study 1 (Names not known) and Study 2 (Names known)

Wave 2 disposition	Study 1 – No names		Study 2 – Names	
	Absolute frequency	Relative frequency	Absolute frequency	Relative frequency
		%		%
<b>No Wave 2 contact made</b>				
Non-working, disconnected	95	11.9	94	9.4
Never answered	17	2.1	45	4.5
<b>Contact made</b>				
Completion	422	52.9	572	57.4
Respondent gone from number	165	20.7	163	16.4
Respondent never available	33	4.1	30	3.0
Respondent refusal/partial	37	4.7	57	5.7
“Gatekeeper” refusal	21	2.6	13	1.3
Incapacitated, deceased	3	0.4	13	1.3
Misc. other	4	0.6	10	1.0
<b>Total</b>	<b>797</b>	<b>100.0</b>	<b>997</b>	<b>100.0</b>

**The nature of the attrition.** As shown in Table 2, the group of Wave 1 respondents in Study 1 who were re-interviewed differed significantly on several factors from those who were lost to the panel. In terms of age, those adults less than 30 years of age at Wave 1 were re-interviewed with only 42% success vs. adults in the 40-59 year group of whom 60% completed Wave 2 surveys. Blacks were significantly less likely to be re-interviewed than Whites. In terms of household income, those respondents who reported Wave 1 annual household incomes of less than \$10,000 were re-interviewed with only 44% success vs. those with incomes over \$20,000 of whom 63% were re-interviewed. Married respondents were more successfully re-interviewed (57%) than those not married (49%). Sixty-two percent of home owners were re-interviewed compared with 47% of renters. The longer one had lived in the neighborhood and the more likely one reported at Wave 1 that he/she would not move, the more likely he/she was re-interviewed.

### 3. STUDY 2

#### 3.1 Methodology

During November and December of 1983, a city-wide (one-stage) RDD survey was conducted by the Northwestern University Survey Laboratory for professors who were examining economic well-being/hardship among Chicago families. (The questionnaire took an average of 20 minutes to administer.) Approximately 3,900 telephone numbers were dialed in the process of completing 997 interviews. For each residence contacted, one head-of-household (male or female) was systematically selected as the designated respondent. Up to 20 call-backs were made to increase the likelihood of completing interviews with hard-to-reach respondents. In total, 1,659 eligible households were reached; those eligibles not interviewed either were unavailable at the time calls were made or refused to participate.

Sixteen months later (Spring 1985), all 997 telephone numbers were re-dialed to gather Wave 2 data. Unlike Study 1, in which respondents were not tracked if they had moved or changed their telephone numbers, an effort was made to find respondents whenever possible, although this effort resulted in only few successes as the respondent's full name (first and last) was typically not available. As in Wave 1 of Study 2, at least 20 call-backs were used with the hardest-to-reach respondents.

More than 80% of respondents had given a name identifier at Wave 1. This information was used by interviewers as follows:

Hello, is this \_\_\_\_\_ ? My name is \_\_\_\_\_ , and I'm calling from Northwestern University. About 16 months ago, in late 1983, we conducted an interview with a (man/woman) named \_\_\_\_\_ at this number. May I please speak with (her/him)?

As with Study 1, the interviewer first verified the telephone number and then gave her/his own name. The third blank contained the pre-recorded name identifier given by the respondent at Wave 1. Those respondents who did not give a name at Wave 1 were asked for by using the same procedure used in Study 1 (*i.e.* asking by gender and age, or by gender only).

#### 3.2 Results

**The magnitude of the attrition.** As shown in Table 1, nearly six in 10 of the Wave 1 sample were re-interviewed (57.4%). Overall, the pattern of Wave 2 dispositions in Study 2 was very similar to what was observed in Wave 2 of Study 1. Of the 425 respondents who were "lost"

**Table 2**  
Respondent Characteristics of Wave 2 Re-interviews for Study 1 (Names not known)  
and Study 2 (Names known)

Respondent characteristic	Percentage Re-interviewed at Wave 2	
	Study 1	Study 2
Gender:		
Females	55	56
Males	49	60
Age:		
< 30 years	42*	--
30-39 years	54	--
40-59 years	60	--
> 59 years	55	--
< 34 years	--	54*
35-49 years	--	64
50-64 years	--	61
> 64 years	--	50
Race:		
Asian	68***	42*
Black	49	53
Hispanic	44	62
White	58	61
Education:		
Not high school graduate	45**	49**
High school graduate	57	58
Some college	50	53
College graduate	60	64
Graduate school	--	68
Household income:		
< \$10,000	44**	--
\$10,000-\$19,999	55	--
\$20,000-\$29,999	63	--
\$30,000 or more	63	--
< \$12,000	--	50**
\$12,000-\$17,999	--	59
\$18,000-\$23,999	--	66
\$24,000 or more	--	64
Marital status:		
Married	57*	57*
Divorced	--	68
Separated	--	43
Single	--	57
Widowed	--	51
Not married	49	--
Residential status:		
Own	62***	59
Rent	47	56
Residential tenure in neighborhood:		
< 3 years	44***	--
3-9 years	55	--
10 or more years	56	--
Likelihood of moving in next 2 years:		
Definitely will	37***	--
Probably will	47	--
Probably will not	57	--
Definitely will not	60	--

Note: Chi-square tests of significance were employed.  
\*\*\*  $p < .001$   
\*\*  $p < .01$   
\*  $p < .05$



to the panel, the greatest proportion (nearly four in 10) was associated with numbers that rang in a new household or in an original household from which the respondent had moved with no new number available. Second most frequent were those persons whose Wave 1 telephone number was no longer in service, accounting for nearly one in four of those lost to the panel. Next in frequency were respondents who refused in some way. The fourth most prevalent reason for losing Wave 1 respondents were those persons whose original telephone numbers were never answered at Wave 2.

**The nature of the attrition.** As shown in Table 2, the group that was interviewed at Wave 2 of Study 2 differed significantly on several factors from the group which was lost, with patterns similar to what was observed in Study 1. In terms of age, those adults less than 34 years of age and those more than 64 years of age at Wave 1 were least likely to be re-interviewed. Asians and Blacks were less likely to be re-interviewed than were Hispanics and Whites. In terms of education, those with less formal education were least likely to be re-interviewed. Those respondents who reported Wave 1 annual household incomes of less than \$12,000 were re-interviewed with only 50% success vs. those with incomes over \$24,000 of whom 64% were re-interviewed. Divorced respondents were most successfully re-interviewed (68%), whereas those who said they were separated at Wave 1 were least likely to be re-interviewed (43%).

## 4. DISCUSSION

### 4.1 Summary of Findings

The two independent studies reported here were two-wave RDD telephone surveys, one with a 12 month lag between waves and the other with a 16 month lag. In Study 1, where names were not known for use at Wave 2, attrition was 47.1%. In Study 2, where name identifiers from Wave 1 were known for 83% of the respondents, attrition was somewhat less, at 42.6%.

This marginal difference in attrition rates ( $\chi^2(1) = 3.51, p < .10$ ) is best considered within the following contextual differences between the studies: Study 1 respondents were **not** explicitly told at Wave 1 that they would be called back a year later and, thus, their names were not asked. In Study 2, respondents were told that they would be called back at some future time. Given the particular nature of the research in Study 1, no effort was made to track Wave 2 respondents who had moved or changed their telephone number. On the other hand, an effort was made to do this in Study 2, although with little success. Study 1 employed a Spanish-language version of the questionnaire; in Study 2, Hispanics who could not speak English were not interviewed.

In both studies, the vast majority of those lost to the panel were respondents who could not be reached via their Wave 1 telephone number, either because the number reached an entirely new residence, the respondent had moved from the household, or the number was no longer in service.

Taken together, the findings of these two telephone studies are fairly consistent with past findings from in-person surveys (e.g. Sobol 1959) in identifying the types of persons most likely to be lost in panel studies. In both Study 1 and 2, younger and older adults, non-Whites, the less educated, and those with lower income were less likely to be re-interviewed than other demographic subgroups.

### 4.2 Implications

Given the cost/benefit attraction of RDD surveys, added to the analytic benefits associated with panel studies, it is worthwhile to consider options that may improve the representativeness

of the final panel in surveys that use RDD for Wave 1 sampling. But before discussing these considerations, the issue of asking for respondents' names in telephone surveys merits further discussion.

**Asking for respondents' names at Wave 1.** As mentioned above, the issue is purportedly one of increasing the likelihood of reaching and, thus, re-interviewing the respondent at Wave 2 vs. the possibility of creating an evaluation apprehension effect (Crano and Brewer 1973) which may bias Wave 2 data. Yet, more than this trade-off enters into consideration.

The issues of confidentiality and informed consent also come into play: it is common practice in academic survey research for a survey organization to never provide respondent telephone numbers to anyone, with the possible exception of the sponsor, and only when he/she is planning a panel study or conducting follow-up interviews with respondents who have explicitly given permission for this. This practice follows from the reasoning that an assurance of confidentiality given to Wave 1 respondents is not violated when respondents are called back as part of the *same* on-going research. The fact that so few Wave 1 respondents refuse to participate at Wave 2, coupled with the observation that it is demographically predictable who is most likely to refuse at Wave 2, provides strong support for the conclusion that calling respondents back without having asked their permission at Wave 1 is *not* a problem.

When a telephone survey sponsor can pay for the expense of tracking respondents who have moved, it appears logical to record respondents' full names at Wave 1, since those who have moved may be tracked through telephone directories; calling new numbers given by telephone company recordings; or, even by calling former neighbors to get a forwarding telephone number in those cases where a respondent's address is also known and a reverse-telephone directory is used. But if respondents will not be tracked at Wave 2, how useful is it to be able to ask for the respondent by name?

It cannot be denied that interviewers say they prefer it. That is, most interviewers feel more comfortable asking for "John" or "John Smith" vs. asking for "a man in his mid-50s." Yet the marginal difference in attrition rates in the two studies reported here, even considering the four-month longer lag time between waves in Study 2 which gathered name identifiers at Wave 1, does not provide compelling evidence of the advantage of names. We acknowledge that an unfortunate limitation of our paper is that other differences in these two RDD panel surveys may have contributed to the observed differential in attrition rates: *e.g.* Wave 2 call-backs were greater in Study 2 (eight in Wave 2 of Study 1 vs. 20 in Wave 2 of Study 2). Thus, this issue will remain unresolved until more controlled research is conducted.

Given the current state of knowledge, we believe that it remains the responsibility of the individual researcher using an RDD panel to weigh the competing tensions of possibly biasing measures of the phenomenon under investigation by alerting respondents that they will be "measured" again (*i.e.* the "reactivity" effect) vs. the possibility of experiencing slightly less attrition by asking for names at the time of the Wave 1 interview.

**Considerations to minimize attrition effects.** Some suggestions can be considered in the attempt to minimize the effects of RDD panel attrition.

Sobol (1959) suggested the possibility of a Wave 1 *over-sampling* of those types of respondents who were most likely to be lost in subsequent waves. At first, this suggestion may sound appealing. This initial appeal follows the reasoning that if one knows who is most likely to be lost, then one can project an over-sampling of those groups at Wave 1. As was shown in Sobol's work, and as found in the two studies reported here, one could estimate what types of persons should be over-sampled at Wave 1; *e.g.* older and younger adults. Over-sampling could be accomplished through the use of a screening procedure introduced late in the Wave 1 field period; (although this clearly would increase Wave 1 total survey costs).



Although it is possible to over-sample, is it also desirable? In asking this question, one is ultimately asking whether the resulting panel is more than just an *on-the-surface* demographic match of the population of interest. In other words, is it enough to merely be concerned with getting, for example, the right number (*i.e.* proportion) of senior citizens in the final wave of a panel, or should one also be concerned whether one has the right “mix” of seniors?

This is an empirical question that the present studies cannot answer. Clearly, more research is needed before survey researchers can be more certain whether it is preferable to over-sample at Wave 1 or to “compensate” for attrition through statistical adjustments to subsequent waves of panel data.

Another aspect of the attrition problem is associated with efforts to minimize the loss of those persons whom interviewers are able to re-contact at Wave 2; *i.e.* respondents who refuse or who are “never at home” to complete the Wave 2 interview. This type of loss accounted for 29% of the Study 1 attrition, and 34% in Study 2. What can be done so that interviewers might be more successful at minimizing these losses, other than merely employing traditional interviewer training techniques and making many call-back attempts?

In this age of microcomputers it is quite feasible for interviewers to be given a Wave 1 “profile” of each respondent, so as to be more familiar with the person to be re-interviewed. Care would have to be exercised to avoid creating expectations on the part of interviewers that might bias respondents’ Wave 2 answers. We are not suggesting that the interviewer necessarily use this information in verbatim form to identify the Wave 1 respondent; we believe name, gender and age are adequate for that purpose. But, there may be subtle changes in an interviewer’s verbal behavior that may lead to increased success at re-interviewing when the interviewer has a more detailed idea of “who” the respondent is. This suggestion must await testing before it can be confidently endorsed, but were it to prove effective without introducing bias into the data, it would be relatively easy to do.

Similarly, introductory statements read by interviewers at Wave 2, could be targeted with special appeals to those demographic groups who appear most likely to refuse at Wave 2: in this case we are referring to the elderly, those with less formal education, those with relatively lower income, and especially those who were rated by Wave 1 interviewers as showing little interest and/or cooperation. Here again, a computer could be programmed to generate special Wave 2 introductory spiels based on Wave 1 data about particular respondents.

These appeals must contain incentives for such persons to participate at Wave 2, as they are often persons with the least intrinsic motivation to participate in surveys. When planning for subsequent waves, surveyors should think of “why” such people would want to cooperate and work such reasoning into the interviewers’ introduction for these persons. Such introductions may be lengthy and may even contain some rapport-building questioning. It may even be possible to give the prospective respondent some feedback about Wave 1 findings, without biasing Wave 2 responses. If so, the respondent may regard the re-contact attempt to be more of a “two-way” exchange.

Regardless, computers could be used to generate these special introductions, which in turn would be matched only with those respondents for whom the message is targeted. Again, we have no empirical evidence to cite regarding the efficacy of this suggestion, but we believe it merits consideration and study.

## 5. CONCLUSION

Our findings suggest that attrition in RDD panels when respondent names are unknown is not of such magnitude as to render the surveying technique invalid or impractical. Due to



its nonreactivity, it would certainly appear to be the preferred approach in two-wave RDD panels in which the researcher has *a priori* reason not to want Wave 1 respondents to know they will be re-contacted. These findings also should provide encouragement for those who are thinking about converting an RDD cross-sectional survey into a panel. We hope that this primarily descriptive paper will encourage other survey methodologists to conduct and report the results of more controlled studies that investigate the nature and magnitude of RDD panel attrition, so that eventually, researchers can more confidently implement strategies to reduce the level of attrition. We suggest that this research should be guided by the observation that reductions in the magnitude of RDD panel attrition appear most likely to occur with well-organized surveying in which each respondent is approached as the individual that he/she is.

### ACKNOWLEDGMENTS

The authors would like to thank Professors Fay Cook, Christopher Jencks, Dan Lewis and Dennis Rosenbaum for permission for access to the data sets which were used for the secondary analyses reported in this paper. The authors also appreciate the helpful comments of Professor Peter V. Miller on an earlier version of this manuscript.

### REFERENCES

- BURGESS, R.D. (1989). Major issues and implications of tracing survey respondents. In *Panel Surveys*, (Eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh). New York: John Wiley & Sons.
- CRANO, W.D., and BREWER, M.B. (1973). *Principles of Research in Social Psychology*. New York: McGraw-Hill.
- CRIDER, D. M., WILLETS, F.K., and BEALER, R.C. (1971). Tracking respondents in longitudinal surveys. *Public Opinion Quarterly*, 35, 613-620.
- DROEGE, R.C., and CRAMBERT, A.C. (1965). Follow-up techniques in a large-scale test validation study. *Journal of Applied Psychology*, 49, 253-256.
- FREEMAN, D.S., THORTON, A., and CAMBURN, D. (1980). Maintaining response rates in longitudinal studies. *Sociological Methods & Research*, 9, 87-98.
- HAUSMAN, J.A., and WISE, D.A. (1979). Attrition bias in experimental and panel data: the Gary income maintenance experiment. *Econometrica*, 47, 455-473.
- LAVRAKAS, P.J. (1987). *Telephone Survey Methods: Sampling, Selection and Supervision*. Newbury Park, CA: Sage.
- LEHNEN, R.G., and KOCH, G.G. (1974). Analyzing panel data with uncontrolled attrition. *Public Opinion Quarterly*, 38, 40-56.
- LEPKOWSKI, J.M. (1989). Treatment of wave nonresponse in panel surveys. In *Panel Surveys*, (Eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh). New York: John Wiley & Sons.
- McALLISTER, R.J., GOE, S.J., and BUTLER, E.W. (1973). Tracking respondents in longitudinal surveys: some preliminary considerations. *Public Opinion Quarterly*, 37, 413-416.
- SOBOL, M.G. (1959). Panel mortality and panel bias. *Journal of the American Statistical Association*, 54, 52-68.
- WINER, R.S. (1983). Attrition bias in econometric models estimated with panel data. *Journal of Marketing Research*, 177-186.

## An Exact Test for the Presence of Stable Seasonality With Applications

BRAJENDRA C. SUTRADHAR, ESTELA BEE DAGUM  
and BINYAM SOLOMON<sup>1</sup>

### ABSTRACT

The X-11-ARIMA seasonal adjustment method and the Census X-11 variant use a standard ANOVA- $F$ -test to assess the presence of stable seasonality. This  $F$ -test is applied to a series consisting of estimated seasonals plus irregulars (residuals) which may be (and often are) autocorrelated, thus violating the basic assumption of the  $F$ -test. This limitation has long been known by producers of seasonally adjusted data and the nominal value of the  $F$  statistic has been rarely used as a criterion for seasonal adjustment. Instead, producers of seasonally adjusted data have used rules of thumb, such as,  $F$  equal to or greater than 7. This paper introduces an exact test which takes into account autocorrelated residuals following an SMA process of the  $(0, q) (0, Q)_s$  type. Comparisons of this modified  $F$ -test and the standard ANOVA test of X-11-ARIMA are made for a large number of Canadian socio-economic series.

KEY WORDS: Standard Anova; Autocorrelated residuals; Seasonality.

### 1. INTRODUCTION

In the analysis of social and economic time series, it is traditional to decompose the observed series into four unobserved components, namely the trend, the cycle, the seasonal variations, and the irregulars.

Socio-economic time series are often presented in seasonally adjusted form so that the underlying short-term trend can be more easily analysed and current socio-economic conditions can be assessed. There are several seasonal adjustment methods available which estimate the seasonal component present in a time series, but the Census X-11 variant (Shiskin, Young and Musgrave 1967) and the X-11-ARIMA method (Dagum 1980) are the most widely applied. To identify the presence of stable seasonality in a time series, the X-11-ARIMA method as well as the Census X-11 variant use the results of the usual  $F$ -test in a one-way ANOVA between monthly seasonal variations and the residuals. However, the residuals in this ANOVA are often autocorrelated, so the nominal significance level of the  $F$ -test may not be valid. Aware of this limitation, producers of seasonally adjusted data, do not guide themselves by the nominal significance level of the  $F$ -test for presence of stable seasonality but by some rule of thumb based on empirical knowledge (see *e.g.* Shiskin and Plewes 1978). In fact, implicit in the X-11-ARIMA test for the presence of 'identifiable seasonality' is that the  $F$ -value for stable seasonality should be greater or equal to 7 if moving seasonality is not present.

The testing for stable seasonality (similarly for annual seasonal shifts) can be approached as a test for the significance of certain regression coefficients in a linear model with autocorrelated errors. The traditional Wald test, the likelihood ratio test, and the tests falling within a generalized least squares framework, all run into convergence problems in testing such a linear

<sup>1</sup> Brajendra C. Sutradhar, Department of Mathematics and Statistics, Memorial University of Newfoundland, St. John's, Newfoundland, A1C 5S7; Estela Bee Dagum, Time Series Research and Analysis Division, Statistics Canada, Ottawa, Ontario, K1A 0T6. Binyam Solomon, Directorate of Social and Economic Analysis, National Defence Headquarters, Ottawa, Ontario, K1A 0K2.

model with highly autocorrelated errors (*cf.* Sutradhar and Bartlett 1990). Pierce (1978) constructed an  $F$ -test based on transformed residuals which are approximately white noise. The transformation suggested in Pierce (1978) is equivalent to using the inversion of the error covariance matrix. But, the inverse of the error covariance matrix may not be obtained for highly autocorrelated errors. Recently Sutradhar, MacNeill and Dagum (1991) proposed a modified  $F$ -test, within a linear model framework for testing for the presence of stable seasonality. Their modified  $F$ -test is derived following Sutradhar, MacNeill and Sahrman (1987), and the test accounts for the presence of autocorrelation in the residuals. The test does not require any transformation or any inversion of the error covariance matrix.

Exact tests for testing the null hypothesis that the seasonal pattern changes over time against the alternative that the seasonal pattern is constant have been developed by Franzini and Harvey (1983). Unlike Franzini and Harvey, the present approach assumes that the seasonal pattern is stable over time possibly at different levels (due to annual shifts) and then tests for the presence of significant stable seasonality.

In most empirical cases, a seasonal moving average (SMA) error model of the  $(0,q)(0,Q)_s$  type is sufficient. In this investigation we simplify the exact test proposed by Sutradhar, MacNeill and Dagum (1991), for such error models. The test is applied to examine for the presence of stable seasonality as well as of annual seasonal shifts in a number of socio-economic series.

The plan of this paper is as follows. Section 2 presents the exact test. Section 3 analyses the results from the application of the modified  $F$ -test to a set of socio-economic time series and compares them with the values given by the X-11-ARIMA method. Section 4 gives the conclusions.

## 2. MODIFIED $F$ -TEST

### 2.1 Selected Model

Consider a stationary seasonal time series  $\{Z_t\}$ , given by

$$Z_t = S_t + U_t, \quad (2.1)$$

where  $Z_t$  is the observed series at time  $t$ ,  $S_t$  is the seasonal component, and  $U_t$  the irregulars. If the time series contains a trend, which is most likely, it is assumed that a suitable detrending technique will yield the model (2.1). In the latter case, the detrended series may be obtained from the original series by taking appropriate differences as in ARIMA modelling (Box and Jenkins 1970) or as is traditionally done by statistical agencies which use the X-11-ARIMA method or Census X-11 variant.

Next, suppose there are  $k$  seasons in a year and there are  $kn$  observations in a time series of  $n$  years. Let  $Z\{(i-1)n+j\}$  be the  $j$ th ( $j = 1, \dots, n$ ) observation under the  $i$ th season ( $i = 1, \dots, k$ ) which corresponds to  $Z_t$  in (2.1). We shall denote in similar manner the  $(i,j)$ th components of  $S_t$  and  $U_t$ , for all  $t = 1, \dots, kn$ . Then, the model assumed for  $S_t$  is (*cf.* Sutradhar and MacNeill 1989):

$$S((i-1)n+j) = \mu + \alpha_i + \beta_j, \quad (2.2)$$

with  $\sum_{i=1}^k \alpha_i = 0$ ,  $\sum_{j=1}^n \beta_j = 0$ .

The  $\alpha$ 's and  $\beta$ 's in (2.2) represent, respectively, the stable seasonality and annual seasonal shifts in the seasonal time series. Thus, when testing for the presence of stable seasonality, we test the hypotheses



$$H_0: \alpha_i = 0 \quad \text{vs.} \quad H_1: \alpha_i \neq 0 \text{ for at least one } i; \quad (2.3)$$

and when testing for the presence of annual seasonal shifts, we test the hypotheses

$$H_0: \beta_j = 0 \quad \text{vs.} \quad H_1: \beta_j \neq 0 \text{ for at least one } j. \quad (2.4)$$

Consequently, the rejection of  $H_0$  in (2.3) and (2.4) would indicate that the series contains significant stable seasonality as well as annual seasonal shifts.

Taking into account model (2.2), the model (2.1) can be written as

$$Z^* = X\gamma + U^*, \quad (2.5)$$

where

$$Z^* = [Z(1), \dots, Z(n), Z(n+1), \dots, Z(kn)]',$$

$$U^* = [U(1), \dots, U(n), U(n+1), \dots, U(kn)]',$$

$$\gamma = [\mu, \alpha_1, \dots, \alpha_{k-1}, \alpha_k, \beta_1, \dots, \beta_{n-1}, \beta_n]'$$

and  $X$  is the appropriate  $kn \times (k+n+1)$  design matrix.

## 2.2 Test Statistics

$U^*$  in (2.5) can be represented by seasonal autoregressive moving average (SARMA) stationary process  $(p, q)(P, Q)_s$ . In most empirical cases we found, however, that a  $(0, q)(0, Q)_s$  model is sufficient. Let  $\Sigma^*$  denote the  $kn \times kn$  covariance matrix of  $U^*$ . Naturally,  $\Sigma^*$  will contain  $\theta \equiv (\theta_1, \dots, \theta_q)$  and  $\Theta \equiv (\Theta_1, \dots, \Theta_Q)$ , where  $\theta$  and  $\Theta$ 's are the parameters associated with the SARMA  $(0, q)(0, Q)_s$  process.

For the usual ANOVA model, viz., when the components of  $U^*$  are i.i.d.  $N(0, \sigma^2)$ , one tests the null hypotheses  $\beta_j = 0$ , and  $\alpha_i = 0$  by using the classical  $F$ -statistics  $F_{A1}$  and  $F_{A2}$  respectively, given by

$$F_{A1} = (k-1)Q_1/Q_3, \quad \text{and} \quad F_{A2} = (n-1)Q_2/Q_3,$$

where

$$Q_1 = k \sum_{j=1}^n (\bar{Z}_j - \bar{Z}_{..})^2, \quad Q_2 = n \sum_{i=1}^k (\bar{Z}_i - \bar{Z}_{..})^2,$$

and

$$Q_3 = \sum_{i=1}^k \sum_{j=1}^n (Z_{ij} - \bar{Z}_i - \bar{Z}_j + \bar{Z}_{..})^2$$

with

$$\bar{Z}_i = \sum_{j=1}^n Z_{ij}/n, \quad \bar{Z}_j = \sum_{i=1}^k Z_{ij}/k, \quad \text{and} \quad \bar{Z}_{..} = \sum_{i=1}^k \sum_{j=1}^n Z_{ij}/kn,$$

$Z_{ij}$  being the  $j$ th observation under the  $i$ th season. In the present set-up, however, these statistics are inappropriate for testing the above hypotheses. This is because, the expected values of the sums of squares are affected by the dependence among observations. Also, sums of squares are not mutually independent. For the case when  $U^*$  in (2.5) follow a SARMA  $(0, q)(0, Q)_s$  process, it can be shown that

$$E(Q_1) = k \sum_{j=1}^n \beta_j^2 + \sigma^2(n-1)C_1(\theta, \Theta),$$

$$E(Q_2) = n \sum_{i=1}^k \alpha_i^2 + \sigma^2(k-1)C_2(\theta, \Theta),$$

and

$$E(Q_3) = \sigma^2(k-1)(n-1)C_3(\theta, \Theta),$$

where, for example, for the SARMA  $(0, 1)(0, 1)_{12}$  process,

$$C_1(\theta, \Theta) = (1 + \theta_1^2)(1 + \Theta_1^2) - (\theta_1/6)(1 + \Theta_1^2)(11 - 1/n) + (2\Theta_1/n)(1 + \theta_1^2) \\ + (\theta_1\Theta_1/6)\{1 - 22/n - (n-2)/n(n-1)\},$$

$$C_2(\theta, \Theta) = (1 + \theta_1^2)(1 + \Theta_1^2) - 2(1 - 1/n)\Theta_1(1 + \theta_1^2) \\ + 1/6\{1 + (1 - 1/n)/11\}\theta_1(1 + \Theta_1^2) - (4/11)(1 - 1/n)\theta_1\Theta_1,$$

$$C_3(\theta, \Theta) = (1 + \theta_1^2)(1 + \Theta_1^2) + (2\Theta_1/n)(1 + \theta_1^2) + (\theta_1/6)(1 + \Theta_1^2)(1 - 1/11n) \\ - (\theta_1\Theta_1/6n)[n/11 - 2(n-2)/11(n-1) - 2].$$

Consequently, the null hypotheses  $\beta_j = 0$ , and  $\alpha_i = 0$  may be tested by using the modified  $F$ -statistics  $F_{M1}$  and  $F_{M2}$  respectively, given by

$$F_{M1} = d_1(\hat{\theta}, \hat{\Theta})F_{A1}, \quad (2.6)$$

$$F_{M2} = d_2(\hat{\theta}, \hat{\Theta})F_{A2}, \quad (2.7)$$

(see also Sutradhar, MacNeill and Sahrman 1987, Sutradhar, MacNeill and Dagum 1991), where  $d_1(\theta, \Theta) = C_3(\theta, \Theta)/C_1(\theta, \Theta)$ ,  $d_2(\theta, \Theta) = C_3(\theta, \Theta)/C_2(\theta, \Theta)$ . The modified  $F$ -statistics  $F_{M1}$  and  $F_{M2}$  account for autocorrelation of the residuals.

Notice that in the independence case when  $\theta = 0$ ,  $\Theta = 0$ ,  $C_1(\cdot) = C_2(\cdot) = C_3(\cdot) = 1$ , which is obvious. In that case the problem reduces to testing the hypotheses by using standard ANOVA  $F$ -statistics.

### 2.3 Computation of $p$ -value

A simulation study (*cf.* Sutradhar and Bartlett 1989, Table IV, p. 1587) indicates that for the cases when  $k$  groups are independent, the distribution of the modified  $F$ -statistics for the SMA/ $(0, q)(0, Q)_s$  process, may be approximated by the usual  $F$ -distribution. In general, the  $F$  approximation to the modified  $F$ -statistic would be inappropriate, in particular when  $k$  groups are correlated and  $n$  is small.

In this paper we use the well known Satterthwaite (1946) approximation (cf. Sutradhar, MacNeill and Dagum 1991) to calculate the  $p$ -value, namely,  $P_r(F_{M1} \geq f_{M1})$ , where  $f_{M1}$  is the data based value of  $F_{M1}$ . In order to do it, we first compute the eigenvalues  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_r > 0 = \lambda_{r+1} = \dots = \lambda_s > \lambda_{s+1} \geq \dots \geq \lambda_n$  of

$$\Sigma^{*1/2} [d_1(\theta, \Theta) D_1 - f_{M1} (I_{kn} - D_2)] \Sigma^{*1/2}, \quad (2.8)$$

where  $d_1(\cdot)$  is given in equation (2.6),  $D_1 = R(RR')^{-1}R'$ , with  $R = C(X'X)^{-1}$ ,  $D_2 = X(X'X)^{-1}X'$ ,  $C$  being a suitable matrix obtained by expressing the:  $H_0 : \beta_j = 0$  in the form  $C\gamma = 0$ , where  $\gamma$  is given in model (2.5). In equation (2.8)  $I_{kn}$  is the  $kn \times kn$  identity matrix. Then the Satterthwaite approximation yields

$$P_r(F_{M1} \geq f_{M1}) = P_r[F_{a,b} \geq bd/ac], \quad (2.9)$$

where  $F_{a,b}$  denotes the usual  $F$ -ratio with degrees of freedom  $a$  and  $b$ , with

$$a = \left( \sum_{j=1}^r \lambda_j \right)^2 / \sum_{j=1}^r \lambda_j^2, \quad b = \left( \sum_{j=s+1}^n \lambda_j \right)^2 / \sum_{j=s+1}^n \lambda_j^2.$$

In equation (2.9),

$$c = \sum_{j=1}^r \lambda_j^2 / \sum_{j=1}^r \lambda_j, \quad d = \sum_{j=s+1}^n \lambda_j^2 / \sum_{j=s+1}^n |\lambda_j|.$$

Similarly,  $P_r(F_{M2} \geq f_{M2})$  may be calculated by using  $d_2(\cdot)$  and  $f_{M2}$  in place of  $d_1(\cdot)$  and  $f_{M1}$  respectively in equation (2.9). The construction of  $D_1$  will now depend on a different  $C$  matrix which will be obtained by expressing the  $H_0 : \alpha_i = 0$  in the form  $C\gamma = 0$ .

### 3. APPLICATIONS

#### 3.1 Monthly Series

The modified  $F$  statistics  $F_{M1}$  and  $F_{M2}$  of equations (2.6) and (2.7) were calculated for a set of 26 monthly series obtained from various economic sectors, namely, Imports, Exports, Consumer Prices and Labour. All series cover the period January 1979 till December 1988 inclusive.

Since the modified  $F$ -test is not valid when moving seasonality is present (except for annual seasonal shifts), none of the series selected are affected by moving seasonality according to certain preliminary tests available in X-11-ARIMA. (We also looked at the plots of the seasonal-irregular ratios.)

The X-11-ARIMA method was applied to obtain the detrended series  $\{Z_t : t = 1, \dots, 120\}$ . Diagnostic checks show that the errors of the detrended series,  $U_t$  (see equation 2.1) follow a  $(0,1)(0,1)_{12}$  SMA model for each of the monthly series. The estimates  $\hat{\theta}_1$  and  $\hat{\Theta}_1$  are used to compute the modified  $F$ -statistics  $F_{M1}$  and  $F_{M2}$ .

In testing for the presence of annual seasonal shifts, the  $p$ -values for the modified  $F$ -test based on the Satterthwaite approximation and on the standard ANOVA  $F$ -test generally were found to be different. For both cases, however, the  $p$ -values were very large for each of the series indicating that there is no moving seasonality in the form of annual shifts.



Table 1  
Diagnostics of Stable Seasonality in Monthly Series

Series	Parameter Estimates		X-11-ARIMA <i>F</i> -Test <sup>a</sup>	Modified <i>F</i> <i>F</i> <sub><i>M</i>2</sub> ( <i>p</i> -value in %)	Final Diagnostic <sup>c</sup>
	$\theta_1$	$\Theta_1$			
IMPORTS					
1. Fodder and feed	-0.09*	-0.01	3.68	3.43(0.06)	Y
2. Coal related materials	0.02	-0.01	64.40	58.76(0.00)	Y
3. Crude vegetable products	0.02	-0.07*	3.48	2.94(0.27)	Y
4. Wool & man made materials	0.02	0.29*	10.98	20.63(0.00)	Y
5. Precious metals	0.27*	0.01	1.25	1.20(31.10)	N
6. Oils & fats	0.41*	0.01	8.59	8.22(0.00)	Y
7. Non-metal minerals	0.04	0.02	16.50	16.68(0.00)	Y
8. Aircraft engines	0.32*	0.00	2.53 <sup>b</sup>	2.36(1.79)	N
9. Other trans. equipments	0.19*	-0.18*	3.48 <sup>b</sup>	2.43(1.31)	N
EXPORTS					
10. Wheat	0.04	-0.03	1.89	1.71(8.71)	N
11. Asbestos	0.13*	-0.03	6.83	6.15(0.00)	Y
12. Wood pulp	-0.27	0.20*	6.45	9.61(0.00)	Y
13. Textile fabrics	0.52*	0.13*	12.05	15.06(0.00)	Y
14. Other fabrics	0.04	0.11*	5.03	6.19(0.00)	Y
15. Television & telecommunication	0.12*	0.01	9.26	8.99(0.00)	Y
16. Domestic export pass.	-0.30*	-0.14*	24.50	18.52(0.00)	Y
CPI					
17. Eggs	-0.04	-0.01	6.90	6.50(0.00)	Y
18. Pasta	-0.05*	-0.04	3.69	3.24(0.10)	Y
19. Onions	-0.42*	-0.03	26.90	23.49(0.00)	Y
20. Housing	0.11*	-0.34*	19.02	9.28(0.00)	Y
21. Clothing	0.03	-0.42*	47.42	24.30(0.00)	Y
22. Transport	-0.09*	-0.02	4.21	3.74(0.02)	Y
LABOUR					
23. Sask. employment (25-34)	-0.19*	-0.11*	67.40	52.35(0.00)	Y
24. Sask. not in labour force	0.12*	-0.36*	22.98	12.69(0.00)	Y
25. Ontario unemployment (25-44)	-0.21*	0.07*	31.4	34.23(0.00)	Y
26. Ontario unemployment male & female (20-24)	-0.02	0.19*	24.27	34.78(0.00)	Y

<sup>a</sup> Critical value is  $F(11,99; 0.01) = 2.47$ .  
<sup>b</sup> X-11-ARIMA and Modified *F* give conflicting inference.  
<sup>c</sup> Y (Yes) – stable seasonality is significant  
N (No) – stable seasonality is not present.  
\* Significant values at 5% level.

To test for the presence of stable seasonality, we computed the  $p$ -values of the modified  $F$ -statistic  $F_{M2}$  (2.7) by using the Satterthwaite approximation and compared them to those given by the X-11-ARIMA  $F$ -test (which is equivalent to the standard ANOVA  $F_{A2}$ ) for the 26 monthly series. The results are shown in Table 1.

The  $p$ -values of the modified  $F$ -statistic in Table 1 show that among the nine import series, three series do not have significant stable seasonality at the 1% significance level (the critical value of  $F(11,99; 0.01) = 2.47$ ). Among the seven exports series, only one series, namely Wheat, appears to have no seasonality. All six CPI series have significant stable seasonality and similarly the four Labour series.

The X-11-ARIMA  $F$ -test values give same results (either rejection or acceptance of the null hypothesis) as the modified  $F$ -test for a large number of series. It seems that for most of the monthly series, under the SMA  $(0,1)(0,1)_s$  error structure, the X-11-ARIMA  $F$ -test (or equivalently standard ANOVA  $F$ -test) is more affected by large negative values of  $\theta_1$ , *i.e.* when there is seasonal autocorrelation in the residuals. This can be generalized by looking at the values of  $C_3(\theta, \theta)/C_2(\theta, \theta)$ . By examining when this fraction is greater or less than 1, it may be seen that the direction of the inequality is affected by the signs of  $\theta_1$  and the size by the value of  $\theta_1$ . Only two series, namely, Imports Aircraft Engines and Imports other transportation Equipments, have standard  $F$ -test values which lead to contradictory conclusions with respect to the modified  $F$ -test. On the other hand, if we would follow the rule of thumb of  $F \geq 7$  to justify seasonal adjustment, then the modified  $F$ -test would be in contradiction for eight out of twelve series. We then seasonally adjusted these eight series with the X-11-ARIMA method and found that the quality of the adjustment was acceptable for six out of the eight cases. All series passed the extrapolation ARIMA model automatically chosen for the program, six out of the eight series passed the X-11-ARIMA guidelines criteria for acceptance; and the four series for which the  $F_{M2}$  values were relatively small, that is, falling between 3.24 and 3.74 were really strongly affected by trading-day variations. Only Imports Fodder and Feed and Imports Crude Vegetable products gave a seasonally adjusted output that could not be considered reliable.

### 3.2 Quarterly Series

The X-11-ARIMA method was applied to four quarterly series of the System of National Accounts to obtain the detrended values ( $z_t, t = 1, \dots, 40$ ). It was found that for all four series  $U_t$  follow a  $(0,1)(0,1)_4$  model. The computation for the modified  $F$ -test is quite similar to the case for monthly series but since the covariance matrix  $\Sigma^*$  is different, the formulas for  $C_1(\cdot)$ ,  $C_2(\cdot)$ , and  $C_3(\cdot)$  in equations (2.6) and (2.7) were adjusted accordingly.

Similar to the monthly series, the  $p$ -values for testing the presence of annual shifts based on the  $F_{M1}$  test were found very large and thus rejecting this pattern of moving seasonality.

The results of the modified  $F_{M2}$  test and the X-11-ARIMA  $F$ -test for testing for the presence of stable seasonality in each of the four series, are given in Table 2. The  $p$ -value for two series namely, Deposits in other Institutions and Small Mortgages are not significant and in agreement with those obtained from X-11-ARIMA. Thus we conclude that these two series contain significant stable seasonality. For the remaining two quarterly series, the modified  $F$ -test and the X-11-ARIMA  $F$ -test give conflicting inferences. Contrary to the X-11-ARIMA  $F$ -test, the modified  $F$ -test yields significant  $p$ -values for these two series. Thus we conclude that these two quarterly series, namely, Net Financial Investments and Corporate claims should not be seasonally adjusted.

**Table 2**  
Diagnostics of Stable Seasonality in Quarterly Series

Series	Parameter Estimates		X-11-ARIMA $F$ -Test <sup>a</sup>	Modified $F$ $F_{M2}$ ( $p$ -value in %) )	Final Diagnostic <sup>c</sup>
	$\theta_1$	$\Theta_1$			
1. Deposits in other institutions	0.53*	0.11*	9.03	9.67(0.04)	Y
2. Net financial investments	0.77*	-0.37*	4.86 <sup>b</sup>	2.56(8.16)	N
3. Small mortgages	0.17*	-0.01	6.65	4.88(1.02)	Y
4. Corporate claims	0.77*	-0.31*	7.88 <sup>b</sup>	3.58(3.20)	N

<sup>a</sup> Critical value is  $F(3,27; 0.01) = 4.51$ .

<sup>b</sup> X-11-ARIMA and Modified  $F$  give conflicting inference.

<sup>c</sup> Y (Yes) - Stable seasonality is significant.

N (No) - Stable seasonality is not present.

\* Significant values at 5% level.

#### 4. CONCLUSIONS

This paper has introduced an exact test for the presence of stable seasonality and annual seasonal shifts based on the modified  $F$ -test by Sutradhar, MacNeill and Sahrman (1987). The new test takes into account the possibility of autocorrelated residuals in the seasonal-irregular ratios of the X-11-ARIMA method. The residuals are assumed to follow a simple Seasonal Moving Average (SMA) model  $(0,q)(0,Q)_s$ . This test is applied to a set of quarterly and monthly series from the system of National Accounts, Imports, Exports, Consumer Prices and Labour. The residuals from the X-11-ARIMA method are found to follow seasonal moving average models (SMA) where either  $\hat{\theta}$  and/or  $\hat{\Theta}$  were significant. The exact  $F$ -test gives values very different from those of the  $F$ -test in X-11-ARIMA (also in the Census X-11 variant) when the autocorrelation of the residuals is of a seasonal character, *i.e.*, whenever  $\hat{\Theta}$  is significantly different from zero.

Among the 26 monthly series analysed, only in two cases, the standard  $F$ -test values gave conflicting conclusions with respect to the modified  $F$ -test. On the other hand, if we would follow the common rule of thumb of  $F \geq 7$  to justify seasonal adjustment, then the modified  $F$ -test gave contradictory results for eight out of twelve series.

By looking at the seasonal adjustment output of these eight series we found that six can be soundly seasonally adjusted by the X-11-ARIMA method.

Concerning the quarterly series, the modified  $F$ -test indicates that there is no stable seasonality in two out of the four series analysed. Furthermore, in one case, the  $F$ -test of X-11-ARIMA gives an  $F$  value greater than 7 whereas the modified  $F$  accepts the null hypothesis.

It has been assumed throughout the paper that moving seasonality may be present in the series only in the form of annual shifts. The present test is not suitable to detect other types of moving seasonal patterns in the series. This raises the necessity of further investigations in this direction.

#### ACKNOWLEDGEMENTS

We would like to thank an anonymous referee for his valuable comments to an earlier version of this paper.



## REFERENCES

- BOX, G.E.P., and JENKINS, G.M. (1970). *Time Series Analysis, Forecasting and Control*. San Francisco: Holden-Day.
- DAGUM, E.B. (1980). *The X-11-ARIMA Seasonal Adjustment Method*. Catalogue 12-564E, Statistics Canada.
- FRANZINI, L., and HARVEY, A.C. (1983). Testing for deterministic trend and seasonal components in time series models. *Biometrika*, 70, 673-682.
- PIERCE, D.A. (1978). Seasonal adjustment when both deterministic and stochastic seasonality are present. In *Seasonal Analysis of Economic Time Series*, (Ed. A. Zellner). Washington, D.C.: U.S. Bureau of the Census, 242-272.
- SATTERTHWAITE, F.E. (1946). An approximate distribution of estimates of variance components. *Biometrics*, 2, 110-114.
- SHISKIN, J., YOUNG, A.H., and MUSGRAVE, J.C. (1967). The X-11 Variant of Census Method II: Seasonal Adjustment Program. Technical Paper 15, Bureau of the Census, U.S. Dept. of Commerce.
- SHISKIN, J., and PLEWES, T. (1978). Seasonal adjustment of the U.S. unemployment rate. *The Statistician*, 27, 181-202.
- SUTRADHAR, B.C., and BARTLETT, R.F. (1989). An approximation to the distribution of the ratio of two general quadratic forms with application to time series valued designs. *Communications in Statistics - Theory and Methods*, 18, 1563-1588.
- SUTRADHAR, B.C., and BARTLETT, R.F. (1990). An Exact Large and Small Sample Comparison of Wald's, Likelihood Ratio and Rao's Tests for Testing Linear Regression with Autocorrelated Errors. Technical Report, Department of Mathematics and Statistics, Memorial University of Newfoundland.
- SUTRADHAR, B.C., and MacNEILL, I.B. (1989). Two-way analysis of variance for stationary periodic time series. *International Statistical Review*, 57, 169-182.
- SUTRADHAR, B.C., MacNEILL, I.B., and DAGUM, E.B. (1991). A Simple Test for Stable Seasonalities. Statistics Canada, Methodology Branch, Working Paper No. TSRA-91-007.
- SUTRADHAR, B.C., MacNEILL, I.B., and SAHRMANN, H.F. (1987). Time series valued experimental designs: One-Way Analysis of Variance with Autocorrelated Errors. In *Time Series and Econometric Modelling*, (Eds. I.B. MacNeill and G.J. Umphrey) Dordrecht: Reidel, 113-129.



# A Theory of Quota Surveys

JEAN-CLAUDE DEVILLE<sup>1</sup>

## ABSTRACT

Simple or marginal quota surveys are analyzed using two methods: (1) behaviour modelling (super-population model) and prediction estimation, and (2) sample modelling (simple restricted random sampling) and estimation derived from the sample distribution. In both cases the limitations of the theory used to establish the variance formulas and estimates when measuring totals are described. An extension of the quota method (non-proportional quotas) is also briefly described and analyzed. In some cases, this may provide a very significant improvement in survey precision. The advantages of the quota method are compared with those of random sampling. The latter remains indispensable in the case of large scale surveys within the framework of Official Statistics.

**KEY WORDS:** Quota surveys; Super-population models; Restricted sampling; Regression estimation.

## 1. INTRODUCTION

Quota sampling is the method most frequently used in France by private polling institutions. It is easy to implement, inexpensive, and has many practical advantages. However, its disadvantages are also well known: likelihood of bias, no possibility of processing non-responses, and the need for external information in order to set the quotas. In the English literature (Cochran 1977; or Madow *et al.* 1983, for example) quotas have a very bad reputation due to the lack of a reliable theory on which statistical inference can be based. The only “defenders” of the method (Smith 1983, in particular) base their arguments on the principles of inference conditional upon sampling, where the sampling plan may generally be ignored.

This paper proposes a theory of quota surveys based on two types of modelling: population behaviour modelling (which is the approach of Smith or the ideas expressed in Gourieroux 1981), and modelling the method of sample collection, which may correspond to a more realistic idea.

In both cases, variance estimates are obtained by resorting to variations of regression estimators.

The first section of the paper describes the quota method and the results of the survey theory that can be subsequently useful. Parts 2 and 3 develop models for the behaviour of individuals in a population, or of those conducting the survey, which justify the method. The last section examines the problems raised, and attempts to demonstrate how the quota method can be used to add to the traditional probabilistic methods, rather than compete with them.

## 2. A BRIEF REVIEW OF THE QUOTA METHOD AND SURVEY THEORY

### 2.1 Cell Quotas; Quotas on the Margins of a Contingency Table – Some Practical Aspects of the Method

At the simplest level, the quota method resembles stratified sampling. The distribution in the population of a discrete characteristic  $h$  possessed by  $N_h$  individuals ( $h = 1$  to  $H$ ) is known.

<sup>1</sup> Jean-Claude Deville, Institut National de la Statistique et des Études Économiques, 18, Boulevard Adolphe Pinard, 75675, Paris Cedex 14, France.



The sample includes  $n_h$  individuals in category  $h$ ; however, the choice of these individuals is left up to the those conducting the survey. The sampling rate  $f_h = n_h/N_h$  may vary from category to category.

In practice, we prefer to control several criteria expressed as  $i, j, \dots, h$  ( $i = 1$  to  $I$ ,  $j = 1$  to  $J$ ,  $\dots$ ,  $h = 1$  to  $H$ ). Ideally, knowing the  $N_{ij\dots h}$  values of the multiple-entry contingency table allows the use of the previous method to define the number  $n_{ij\dots h}$  of members in the sample depending upon the  $f_{ij\dots h}$  rates. Except in very specific cases (few criteria having few modalities each) this method is unrealistic, because it leads to a search for individuals who are extremely difficult to find.

Thus, it is preferable to use **marginal quotas**, by calibrating the sample so that its distribution in accordance with the first criterion leads to a given  $n_{i+ \dots +}$  number of members, and the same is done for the other criteria. The only constraint on these marginal values is that they must be added to the overall sample size  $n$ . However, in practice, a single sampling rate  $f$  is adopted for each set of quotas:  $n_{i+ \dots +} = fN_{i+ \dots +}$ ,  $n_{+j \dots +} = fN_{+j \dots +}$  and  $n_{+ \dots +h} = fN_{+ \dots +h}$  with the obvious notations (+ in place of an index indicates the addition of all the modalities in the category represented by the index).

Beyond the obvious collection advantages, this technique is the one most often imposed by the external data on which the quotas are based. These are obtained, for example, from various sources, thus preventing any cross-correlations. Another situation arises when the quotas are established on the basis of a large survey (a labour survey, for example): each distribution is done in accordance with a criterion (age, socio-professional category, *etc.*) that may be considered to be reliable. On the other hand, the cross-correlations are affected by a large random error, and cannot be used to set the quotas.

In practice, the quota method is most often used to complement more traditional methods as the last sampling technique used in a multi-stage stratified survey on a geographic basis (region, size of the agglomerations). Each primary unit is assigned to a survey officer for whom quotas have been set. The survey officer also receives instructions to distribute his sample in order to make data collection as close to random as possible.

## 2.2 Traditional Survey Theory

We want to measure the total  $Y$  of a variable whose value  $Y_k$  for individual  $k$  is fixed, with no randomness. Only sample  $s$  is random, and the law of probability that governs  $s$  is known, since it is controlled by the statistician. Thus, we also know the possibility  $\pi_k$  that each individual will appear in  $s$ . Without any other information, the natural (unbiased) estimator to be used is the estimator based on inflated values:

$$\hat{Y} = \sum_{k \in s} Y_k / \pi_k = \sum_s d_k Y_k \quad \text{with} \quad d_k = 1 / \pi_k.$$

When the  $\pi_k$  are all equal to  $n/N$ , the sampling rate, we have:

$$\hat{Y} = N/n \sum_s Y_k = N\bar{y},$$

where  $\bar{y}$  is the mean of  $Y$  in the sample.

This estimator has a known variance, which is a quadratic form  $V(Y_U)$  on the vector of  $Y_k$  in the population:

$$\text{Var}(\hat{Y}) = V(Y_U) = \sum_k Y_k(d_k - 1) + \sum_{kl} Y_k Y_l d_k d_l (\pi_{kl} - \pi_k \pi_l), \quad (2.2.1)$$

where  $\pi_{kl}$  is the probability of simultaneously having  $k$  and  $l$  in  $s$ .

Similarly, the variance of  $\hat{Y}$  can be estimated by a quadratic form on vector  $Y_s$  of the  $Y_k$  in the sample:

$$\hat{V}(Y_s) = \sum_{kl \in s} \Delta_{kl} Y_k Y_l,$$

with

$$\begin{aligned} \Delta_{kl} &= (1 - \pi_k) / \pi_k^2 \quad \text{if } k = l \\ &= (\pi_{kl} - \pi_k \pi_l) / (\pi_k \pi_l) \quad \text{if } k \neq l. \end{aligned}$$

Depending upon the sampling plans, these expressions take the specific forms found in the manuals (Desabie 1965; Cochran 1977; Wolter 1985).

Any external information can improve the quality of the estimate. This is usually presented in the form of a vector  $X$  in which each of the  $p$  components is the total of a measurable variable in each of the possible samples. The estimate of  $Y$  can thus be improved by using regression estimation:

$$\hat{Y}_{\text{Reg}} = \hat{Y} + (X - \hat{X})' \hat{B},$$

where  $B$  is the vector of the coefficients of the regression of the  $Y_k$  on the  $X_k$  estimated by:

$$\hat{B} = \sum_s (d_k X_k X_k')^{-1} \sum_s d_k X_k Y_k.$$

When the constant is part of the regressors, or if it is a linear combination of the regressors and the sample has equal probabilities, the formula is simplified as follows:

$$\hat{Y}_{\text{Reg}} = X' \hat{B}.$$

The variance of  $\hat{Y}_{\text{Reg}}$  is simply expressed by introducing the residuals of the regression  $E_k = Y_k - X_k' \hat{B}$  into the population. We know that we have:

$$\text{Var}(\hat{Y}_{\text{Reg}}) = V(E_U)$$

thus, we introduce in formula (2.2.1) vector  $E_U$  of residuals  $E_k$ . At the same time, we approximate an estimate of this variance by  $\hat{V}(e_s)$ , where  $e_s$  is the vector of  $e_k = Y_k - X_k' \hat{B}$ , the estimated residuals of the regression.

Under some sampling plans, these expressions assume particular forms. As a general rule,  $V$  and  $\hat{V}$  are the positive quadratic forms, and the  $E_k$  or  $e_k$  quantities smaller than the  $Y_k$ ; the regression estimator leads to substantial improvements over the inflated values.

A particularly important case that we will use later is one where  $X$  is a vector of the total accounting variables (values on the basis of which the quotas are constructed). Typically, the additional information is the vector of dimension  $I + (J - 1) + \dots + (H - 1)$  formed by the quantities:  $N_{i+...+}, N_{+j+...+}, N_{+...+h}$  for  $i = 1$  to  $I$ ,  $j = 1$  to  $J - 1$ , and  $h = 1$  to  $H - 1$  (keeping only those variables that are linearly independent). Thus, the regressors

are the indicative variables of categories  $i$  ( $i = 1$  to  $I$ ),  $j$  ( $j = 1$  to  $J - 1$ ), and  $h = 1$  to  $(H - 1)$ . Since the constant is a linear combination of the regressors (it is the sum of the first  $I$  of them), the regression estimator takes the form:

$$\hat{Y}_{\text{Reg}} = \sum_i N_{i+...+} \hat{A}_i + \sum_j N_{+j...+} \hat{B}_j + \dots + \sum_h N_{++...+h} \hat{C}_h, \quad (2.2.2)$$

where  $\hat{A}_i$  (for example) indicates belonging to category  $i$ .

If we are only working with a single category, the regressors are orthogonal 2 by 2 and we have:

$$\hat{Y}_{\text{Reg}} = \sum_i N_i \hat{Y}_i$$

where  $\hat{Y}_i$  is the estimator of the mean of  $Y$  in category  $i$ . Thus,  $i. \hat{Y}_{\text{Reg}}$  is nothing but the post-stratified estimator.

### 2.3 Sampling Theories Based on Models

In this approach, we consider that the  $Y_k$  are random variables governed by a super-population model. This consists of parameters that we estimate on the basis of the sample. We can then calculate the probability, under the estimated model, of the non-observed values of  $Y$ , that is,  $\hat{Y}_k$ . The prediction estimator is the sum of the observed and predicted values and can be obtained as follows:

$$\hat{Y}_{\text{Pred}} = \sum_s Y_k + \sum_{U-s} \hat{Y}_k.$$

If, for example, in an equal probabilities survey, the model is a regression  $Y_k = X'_k \cdot \beta + \epsilon_k$ ,  $\epsilon_k$ , when the  $k$  values are independent, centred, and of equal variance, and when the constant appears on the regression (or when we have a linear combination of  $X_k$  that is constant), we have:  $\sum_s Y_k = \sum_s X'_k \beta$ ; and the prediction estimator and the regression estimator are the same.

We say that  $\hat{Y}$  is without bias under the model when, for all  $s$ ,  $\mathcal{E}(\hat{Y} - Y) = 0$  (conditionally upon the sample, the probability and variance under the model are expressed as  $\mathcal{E}$  and  $\mathcal{V}$ ). For the prediction estimator, we must only have, for all  $k$ , the natural condition  $\mathcal{E} \hat{Y}_k = \mathcal{E} Y_k$ , in order for this to be true. With the model, we can also evaluate the average quadratic deviation:  $\mathcal{E}(\hat{Y}_{\text{Pred}} - Y)^2$ , since we know that the two terms  $\hat{Y}_{\text{Pred}}$  and  $Y$  are random, and that  $\hat{Y}_{\text{Pred}}$  depends upon sample  $s$ . The above-mentioned probability is thus conditional upon sample  $s$ . This follows a certain probability law already discussed in the previous paragraph. The precision of this estimator can be measured by calculating:

$$\mathcal{V}(\hat{Y}_{\text{Pred}}) = \mathcal{E} \mathcal{E}(\hat{Y}_{\text{Pred}} - Y)^2.$$

If the law of  $s$  is such that the  $Y_k$  are independent (the so-called non-informative sampling), then this quantity equals:

$$\mathcal{E}(\mathcal{E}(\hat{Y}_{\text{Pred}} - Y)^2),$$

where the internal probability is conditional upon  $Y_k$ . If  $\hat{Y}_{\text{Pred}}$  is equal to  $\hat{Y}_{\text{Reg}}$ , and we have a condition of independence, we will have:

$$\mathcal{V}(\hat{Y}_{\text{Pred}}) = \mathcal{E}(\text{Var}(\hat{Y}_{\text{Reg}})).$$



## 2.4 Comments on the Two Approaches Applied to the Quota Method

a) In both cases, the process of estimation will be effective if the variable of interest is well explained by category indicators on which the quotas are roughly based, because the regression adjustment residuals will be small.

b) In a quota survey the “sampling plan” is not known by the statistician. Thus, he cannot make inferences without using a model. The latter may be a population behaviour model (“model” approach) that requires him to assume certain responsibilities regarding the nature of what he observes. This approach will be developed in the second part of this paper. This may also consist of modelling the sampling plan; which means taking responsibility for the operation of the collection process. This approach will be developed in the third section of this paper.

In all cases, the modelling speculation must be mobilized in order to validate a kind of inference. The question is to know whether it is easier and more plausible to model the behaviour of the individuals surveyed, or to model the sample collection process (including the contacts between interviewer and interviewee).

c) In this respect, the hypothesis made in section 2.3 regarding the independence between randomness in the population and randomness in the collection process is **crucial**. If sampling is controlled by the statisticians, this guarantee can be ensured, except for the effect of non-responses. In the case of the quota method, there are no guarantees. Let us assume, for example, that we want to measure incomes  $Y_k$ , the probability  $\pi_k$  of finding  $k$  in the sample may be very low if  $Y_k$  is large. In other words, the fact of belonging to the sample (which is 1 if  $k$  is in  $s$ , and 0 otherwise) and the residual of the super-population model  $\epsilon_k$  are negatively correlated. This example illustrates well the main danger of the quota method, which the following theory does not take into account.

## 3. QUOTA THEORY WITH A SUPER-POPULATION MODEL

### 3.1 Cell Quotas

There is a single cell category  $i = 1$  to  $I$  for the known values  $N_i$ . The model that can be imagined is as follows:

$$Y_k = m_i + \epsilon_k, \quad (3.1.1)$$

$\epsilon_k$  centred independently of variance  $\sigma_i^2$  where  $i$  is the cell to which  $k$  belongs.

The Gauss-Markov estimators of  $m_i$  are the means observed in the various  $\bar{y}_i$  cells. Thus, the prediction estimator is:

$$\hat{Y}_{\text{Pred}} = \sum_i (N_i - n_i) \bar{y}_i + \sum_i n_i \bar{y}_i = \sum_i N_i \bar{y}_i. \quad (3.1.2)$$

This has the form of the post-stratified estimator. Moreover:

$$\text{Var}(\hat{Y}_{\text{Pred}} - Y)^2 = \sum_i \sigma_i^2 N_i (N_i - n_i) / n_i. \quad (3.1.3)$$

This quantity does not depend upon sample  $s$ , as the latter always includes (with a probability of 1 !)  $n_i$  individuals in cell  $i$ .

$E\mathcal{E}(\hat{Y}_{\text{Pred}} - Y)^2$  can be estimated by replacing  $\sigma_i^2$  by its usual estimator  $s_i^2 = (n_i - 1)^{-1} \sum_{k \in s_i} (Y_k - \bar{y}_i)^2$  with  $s_i$  being part of  $s$  in cell  $i$ .

These results are from Gourieroux (1981) and represent, to a certain extent, a justification of the simple quota method.

### 3.2 Marginal Quotas – “Representative” Case

In this and the following paragraphs, we will restrict ourselves to the case of quotas overlapping 2 criteria  $i$  and  $j$ . The generalization with more than 2 criteria does not pose any particular problems, but leads to very complex notations that we prefer to avoid (see Appendix).

Thus, the situation is as follows: the values  $N_{i+}$  and  $N_{+j}$  of the two universe breakdowns are known. The sampling only allows samples of fixed size  $n = fN$  including  $n_{i+} = fN_{i+}$  individuals for each  $i$ , and  $n_{+j} = fN_{+j}$  individuals for each  $j$ .

We postulate an analysis of variance model in the population, formulated as follows:

If  $k$  belongs to cell  $(i, j)$ :

$$Y_k = \alpha_i + \beta_j + \epsilon_k. \quad (3.2.1)$$

The  $\epsilon_k$  are centred, independent, and we have  $\text{Var } \epsilon_k = \sigma_i^2 + \gamma_j^2$ .

For reasons of identification of the model, we postulate that  $\beta_J = 0$ .

This is equivalent to postulating that  $Y_k = (\alpha_i + u_{ik}) + (\beta_j + v_{jk})$  where  $u_{ik}$  and  $v_{jk}$  are independent, and their respective variances are  $\sigma_i^2$  and  $\tau_j^2$ .

We estimate  $\alpha_i$  and  $\beta_j$  using the ordinary least squares (OLS) method, because we ignore the values of the variance elements; the  $\hat{\alpha}_i$  and  $\hat{\beta}_j$  are solutions of the system:

$$\begin{aligned} \sum_j n_{ij} \bar{y}_{ij} &= n_{i+} \hat{\alpha}_i + \sum_j n_{ij} \hat{\beta}_j \quad (i = 1 \text{ to } I) \\ \sum_i n_{ij} \bar{y}_{ij} &= n_{+j} \hat{\beta}_j + \sum_i n_{ij} \hat{\alpha}_i \quad (j = 1 \text{ to } J - 1), \end{aligned} \quad (3.2.2)$$

with  $\bar{y}_{ij}$  the mean of the  $Y_k$  over the  $s_{ij}$  part of the sample in cell  $(i, j)$ . Thus, the prediction estimator can be written as follows:

$$\hat{Y}_{\text{Pred}} = \sum_{ij} (N_{ij} - n_{ij}) (\hat{\alpha}_i + \hat{\beta}_j) + \sum_{ij} n_{ij} \bar{y}_{ij}.$$

**Result 1:** Under model (3.2.1), the prediction estimator using the OLS is  $N\bar{y}$ . We check that it is unbiased for the model; that is, that  $\mathcal{E}(N\bar{y} - Y) = 0$ .

**Proof:** Immediately from (3.2.2), and because of the fact that the quotas are proportional to the numbers in the population.

**Result 2:** We have:

$$\mathcal{E}(N\bar{y} - Y)^2 = (N^2/n)(1 - f)n^{-1} \left( \sum_i n_{i+} \sigma_i^2 + \sum_j n_{+j} \tau_j^2 \right).$$

This quantity does not depend upon the sample (as it depends only upon the quotas). Thus, to a certain extent, this is a justification for the marginal quotas method.

**Proof:** With  $m_k = \varepsilon Y_k$ , using the unbiased character of the estimator we have:

$$\begin{aligned}\varepsilon(N\bar{y} - Y)^2 &= \varepsilon \left( (N/n) \sum_s (Y_k - m_k) - \sum_U (Y_l - m_l) \right)^2 \\ &= \varepsilon \left( (N/n) \sum_s \epsilon_k - \sum_U \epsilon_l \right)^2 \\ &= (N/n)^2 \sum_{ij} n_{ij}(\sigma_i^2 + \tau_j^2) - 2(N/n) \sum_{ij} n_{ij}(\sigma_i^2 + \tau_j^2) + \sum_{ij} N_{ij}(\sigma_i^2 + \tau_j^2).\end{aligned}$$

But

$$\begin{aligned}\sum_{ij} N_{ij}(\sigma_i^2 + \tau_j^2) &= \sum_i N_{i+} \sigma_i^2 + \sum_j N_{+j} \tau_j^2 \\ &= (N/n) \left( \sum_i n_{i+} \sigma_i^2 + \sum_j n_{+j} \tau_j^2 \right)\end{aligned}$$

from which:

$$\begin{aligned}\varepsilon(N\bar{y} - Y)^2 &= (N^2/n) (1 - f) n^{-1} \left( \sum_{ij} n_{ij}(\sigma_i^2 + \tau_j^2) \right) \\ &= (N^2/n) (1 - f) \left( \sum_i p_{i+} \sigma_i^2 + \sum_j p_{+j} \tau_j^2 \right) \\ &\quad \text{with } p_{i+} = N_{i+}/N \text{ and } p_{+j} = N_{+j}/N.\end{aligned}$$

The estimate of the precision of  $E(N\bar{y} - Y)^2$  is derived from this. In fact, with this model,  $s_{ij}^2$  has a probability of  $\sigma_i^2 + \tau_j^2$ . Thus, an unbiased estimator of the precision is obtained by

$$(N/n)^2 (1 - f) \sum_{ij} n_{ij} s_{ij}^2$$

if all the  $n_{ij}$  are equal to or greater than 2.

This estimator is formally identical to the one that we would use in a complete post-stratification on cells  $(i, j)$ . We can also use  $(N/n)^2 (1 - f) \sum_s e_k^2$ , where  $e_k$  are the estimated residuals of the model.

### 3.3 What Happens if the Model is False?

**3.3.1** An initial way of looking at the question is to put model (3.2.1) into the general model where the mean of  $Y_k$  depends upon the pair  $(i, j)$ . This can be written as follows:

$$Y_k = \alpha_i + \beta_j + \gamma_{ij} + \epsilon_k, \quad (3.3.1.1)$$

with the usual hypotheses for  $\epsilon_k$  and the terms of interaction  $\gamma_{ij}$  that verify the constraints of identifiability:

$$\sum_j N_{ij} \gamma_{ij} = 0 \quad \text{and} \quad \sum_i N_{ij} \gamma_{ij} = 0. \quad (3.3.1.2)$$



Thus we have:

$$\mathcal{E}(N\bar{y} - Y) = \sum_{ij} (Nn_{ij}/n - N_{ij})\gamma_{ij}, \quad (3.3.1.3)$$

such that the estimator is biased for the model except when  $n_{ij} = fN_{ij}$ , which has no reason to exist.

This means that the terms of sum (3.3.1.3) may well compensate for each other, since their signs are *a priori* undetermined.

On the other hand, if “good” sampling precautions are taken,  $Nn_{ij}/n - N_{ij}$  should usually be close to 0.

It is clear, in any case, that the more suitable the additive model is (small  $\gamma_{ij}$ ), and the more the sampling plan approaches randomness, the more likely it is that bias will be reduced.

**3.3.2** Another way to view the misrepresentation of the model, which has already been described, is to no longer admit that there is independence between the randomness of the sample and the randomness of the additive model. This means that distinct models should be developed for the  $(Y_k, k \in S)$  and  $(Y_l, l \notin S)$  vectors. This approach has often been used in the econometric literature, to which the reader is referred. It is clear that risk-taking in regards to the data becomes enormous, and is often incompatible with objective work on the part of the statistician.

### 3.4 Marginal Quotas with Unequal Rates

In the case of cell quotas, we can arbitrarily set quotas for each cell. Until now, in the case of marginal quotas, we have only examined the case where the quotas were proportional to the size of the population.

In many cases however, we may be tempted to over-represent certain categories. If, for example, we want to study household assets, we may want to set the largest quotas for older households (quotas by age group), on the one hand; and for those where the head is self-employed (quotas by social categories), on the other.

Thus, we formally force the sample to fall within a given size  $n_{i+}$  and  $n_{+j}$  (however, the sum of  $n_{i+}$  is always equal to the sum of  $n_{+j}$ ).

In this case, always using the OLS as an estimation technique, we can easily find that the total prediction estimator is:

$$\hat{Y}_{\text{Pred}} = \sum_i N_{i+} \hat{\alpha}_i + \sum_j N_{+j} \hat{\beta}_j, \quad (3.4.1)$$

$\hat{\alpha}_i$  and  $\hat{\beta}_j$  always verify estimating equations (3.2.2). It is easy to see that this estimator may be expressed as follows:

$$\hat{Y}_{\text{Pred}} = \sum_{ij} (w_i^{(1)} + w_j^{(2)}) n_{ij} \bar{y}_{ij} = \sum_{ij} \hat{N}_{ij} \bar{y}_{ij}.$$

Thus, the quantities  $(w_i^{(1)} + w_j^{(2)}) n_{ij}$  seem to be estimates of the size of cells  $(i, j)$ , an idea that will be largely exploited in the following sections.

On the other hand, the variance of this estimator under the model depends upon all the  $n_{ij}$ , and this can be demonstrated by a rather cumbersome calculation. The justification of the quota method described above no longer works.

## 4. MODELS FOR THE SAMPLING PLAN

### 4.1 A Model Sampling Plan

The idea is one of a simple random sampling constrained by the quotas imposed. The selection algorithm, while totally unrealistic, consists of drawing a series of simple random samples until we find one that verifies the quotas. Thus, each sample that verifies the quotas has the same positive probability of being drawn, the samples that do not verify the quotas have a zero probability of being drawn.

The purpose is to model the fact that the person conducting the survey will correctly follow the dispersion constraints on the survey units assigned to him.

### 4.2 Cell Quotas

This sampling model is based on an *a priori* stratification. Its practical advantage is that it does not require a sampling frame where the stratification variables are present. It is implemented rigorously in certain cases, for example, in a telephone survey based on a non-informative random list of telephone numbers, and when surveys are carried out only until the quotas are met.

The formulas that provide the estimators, the variances, and the precision estimates are those given in all the manuals. They have a certain similarity with those described in section 3.1 (see Gouriboux 1981).

### 4.3 The Case of Marginal Quotas: General Estimators

The sampling model is that of simple random sampling constrained by marginal quotas. SRS provides samples with  $n_{ij}$  members in the various cells that can be taken as a random vector (in whole values) in  $R^{IJ}$ . The quota constraint means that we are limited to a random vector as follows:

$$\sum_j n_{ij} = n_{i+} \quad (i = 1 \text{ to } I) \quad \text{and} \quad \sum_i n_{ij} = n_{+j} \quad (j = 1 \text{ to } J - 1),$$

that is, one that varies within a sub-space of size  $IJ - I - J + 1$ . We place ourselves in the case where the overall sampling rate is negligible, and the law of the  $n_{ij}$  can be compared to a multinomial law ( $n, p_{ij} = N_{ij}/N$ ).

Conditional upon  $n_{ij}$ , the  $\bar{y}_{ij}$  estimate the  $\bar{Y}_{ij}$  without bias. The idea is now to construct an estimator of the total of  $Y$  by weighting the  $\bar{y}_{ij}$  by the estimators of  $N_{ij}$ , that is, the  $p_{ij}$ . If we choose to maximize the probability, this is proportional to:

$$\prod_{ij} p_{ij}^{n_{ij}}. \quad (4.3.1)$$

Thus, we maximize

$$\sum_{ij} n_{ij} \text{Log} p_{ij} \quad (4.3.2)$$

under the following constraints

$$\sum_j p_{ij} = p_{i+} \quad (i = 1 \text{ to } I) \quad \text{and} \quad \sum_i p_{ij} = p_{+j} \quad (j = 1 \text{ to } J - 1) \quad (4.3.3)$$

which leads to solving the system for  $a_i, b_j$  ( $p_{i+} = N_{i+}/N$ ,  $p_{+j} = N_{+j}/N$  are known):

$$\sum_j \hat{p}_{ij}^{\circ} (a_i + b_j)^{-1} = p_{i+} \quad (i = 1 \text{ to } I) \quad (4.3.4)$$

$$\sum_i \hat{p}_{ij}^{\circ} (a_i + b_j)^{-1} = p_{+j} \quad (j = 1 \text{ to } J - 1; b_J = 0),$$

with  $\hat{p}_{ij}^{\circ} = n_{ij}/n$  frequency in the sample.

The estimators of  $p_{ij}$  are thus  $\hat{p}_{ij}^{\circ} (a_i + b_j)^{-1}$  and the estimator we are looking for can be written as follows:

$$\hat{Y}_Q = (N/n) \sum_{ij} n_{ij} (a_i + b_j)^{-1} \bar{y}_{ij} = (N/n) \sum_s w_k Y_k, \quad (4.3.5)$$

where  $w_k = (a_i + b_j)^{-1}$  is the weight added to  $Y_k$  in the case when  $k$  appears in cell  $(i, j)$ . This estimator is asymptotically without bias under the SRS model in  $U$ , as are the maximum probability estimators. The quotas do not play an explicit role in (3.3.4), but they affect the values of  $a_i$  and  $b_j$ .

In the normal case when the marginal quotas are “proportional”, with a fixed sampling fraction  $f$ , the solution of equations (4.3.4) is evident:  $a_i = 1$  for any  $i$ , and  $b_j = 0$  for any  $j$ . The estimator of the total is  $N\bar{y}$ , as could be expected, and has the same expression as the equal-probability probabilistic sampling.

**Comment:** The use of maximum probability to estimate the proportions is rather arbitrary. A chi-square criterion (minimize  $\sum_{ij} (p_{ij} - \hat{p}_{ij}^{\circ})^2 / \hat{p}_{ij}^{\circ}$ ) would make the (4.3.4) system linear.

#### 4.4 Variance of the Estimator and its Estimate

**4.4.1** To establish a variance formula we will use the parametrization of variable  $Y$  used by J.C. Deville and C.E. Särndal (1990), which we will express in the form of a:

**Lemma:** For any variable  $Y = (Y_k; k \in U)$ , we can choose an uniquely defined parametrization

$$Y_k = \bar{Y}_{ij} + R_k \quad \text{if } k \text{ is in cell } (i, j) \quad (k \in U_{ij}) \quad \text{with} \quad \sum_{k \in U_{ij}} R_k = 0,$$

$$\bar{Y}_{ij} = A_i + B_j + E_{ij} \quad \text{with} \quad B_J = 0$$

$$\sum_j N_{ij} E_{ij} = 0 \quad i = 1 \text{ to } I$$

$$\sum_i N_{ij} E_{ij} = 0 \quad j = 1 \text{ to } J - I.$$

In fact,  $A_i$  and  $B_j$  are numbers that minimize the quantity  $\sum_U (Y_k - A_i - B_j)^2$  where, in an equivalent manner  $\sum_{ij} N_{ij} (\bar{Y}_{ij} - A_i - B_j)^2$ .



Thus, we can write:

$$\hat{Y}_Q = (N/n) \sum_{ij} n_{ij} (a_i + b_j)^{-1} (A_i + B_j + E_{ij} + \bar{R}_{ij}) \quad \text{where} \quad \bar{R}_{ij} = \sum_{s_{ij}} R_k / n_{ij}.$$

Taking into account equation 4.3.4 and the lemma:

$$\hat{Y}_Q - Y = \sum_{ij} \hat{N}_{ij} (E_{ij} + \bar{R}_{ij}) \quad \text{with} \quad \hat{N}_{ij} = (N/n) n_{ij} (a_i + b_j)^{-1}, \quad (4.4.1)$$

which is the basic expression for the calculation of the variance.

Conditional upon  $n_{ij}$ , the  $\hat{N}_{ij}$  are constant, and sub-samples  $s_{ij}$  are independent simple random samplings. Thus we have:

$$\text{Cond bias}(\hat{Y}_Q) = \sum_{ij} \hat{N}_{ij} E_{ij} = N \sum_{ij} \hat{p}_{ij} E_{ij}$$

$$\text{Cond Var}(\hat{Y}_Q) = \sum_{ij} \hat{N}_{ij}^2 V_{ij} / n_{ij} \quad \text{where} \quad V_{ij} = (1/N_{ij}) \sum_{U_{ij}} R_k^2.$$

Thus (demonstration in the Appendix) we have:

**Result 1:**

$$\text{Var} \left( \sum_{ij} \hat{p}_{ij} E_{ij} \right) = 1/n \sum_{ij} p_{ij} E_{ij}^2.$$

Furthermore, the probability of  $\hat{p}_{ij}^\circ (a_i + b_j)^{-1}$  is (in terms close to  $1/n$ )  $p_{ij} (a_i^\circ + b_j^\circ)^{-1}$  where  $a_i^\circ$  and  $b_j^\circ$  are the solutions to equations (4.3.4), in which  $\hat{p}_{ij}^\circ$  are replaced by the exact  $p_{ij}$ .

This leads to:

**Result 2:** The variance of the quota estimator  $\hat{Y}_Q$  is given by:

$$\text{Var}(\hat{Y}_Q) = (N^2/n) \sum_{ij} p_{ij} (E_{ij}^2 + (a_i^\circ + b_j^\circ)^{-1} V_{ij}).$$

If the quotas are proportional to the size of the population, we will have:

$$\text{Var}(\hat{Y}_Q) = (N^2/n) \sum_{ij} p_{ij} (E_{ij}^2 + V_{ij}).$$

#### 4.4.2 Estimating the Variance

The conditional variance of  $\hat{Y}_Q$  can be estimated by:

$$\sum_{ij} \hat{N}_{ij}^2 s_{ij}^2 / n_{ij} = (N^2/n) \sum_{ij} \hat{p}_{ij} (a_i + b_j)^{-1} s_{ij}^2,$$

where  $s_{ij}^2$  is the usual unbiased estimator of  $V_{ij}$ . The probability of the square of the conditional bias is  $(N^2/n) \sum_{ij} p_{ij} E_{ij}^2$  and is estimated by  $(N^2/n) \sum_{ij} \hat{p}_{ij} \hat{E}_{ij}^2$  where  $\hat{E}_{ij} = \bar{y}_{ij} - \hat{A}_i - \hat{B}_j$  and  $\hat{A}_i$  and  $\hat{B}_j$  are the solutions of:

$$\begin{aligned} \sum_j \hat{p}_{ij} (\hat{A}_i + \hat{B}_j) &= \sum_j \hat{p}_{ij} \bar{y}_{ij} \quad (i = 1 \text{ to } I), \\ \sum_j \hat{p}_{ij} (\hat{A}_i + \hat{B}_j) &= \sum_j \hat{p}_{ij} \bar{y}_{ij} \quad (j = 1 \text{ to } J - I) \quad \text{with } B_J = 0. \end{aligned} \quad (4.4.2)$$

In other words, the estimate of  $E_{ij}$  is obtained by fitting to the data an additive ANOVA model without interaction, the fitness criterion being that of least squares weighted by  $(a_i + b_j)^{-1}$ .

Thus, the variance estimator is:

$$\widehat{\text{Var}}(\hat{Y}_Q) = (N^2/n) \sum_{ij} \hat{p}_{ij} (\hat{E}_{ij}^2 + (a_i + b_j)^{-1} s_{ij}^2). \quad (4.4.3)$$

When the quotas are proportional to the population numbers, this expression can be simplified as follows:

$$(N^2/n) \sum_{ij} n_{ij} (\hat{E}_{ij}^2 + s_{ij}^2)/n. \quad (4.4.4)$$

If the  $n_{ij}$  are all sufficiently large that  $n_{ij}/(n_{ij} - 1) = 1$ , the sum of the formula is the sum of the squares of the residuals estimated in the OLS adjustment of the  $Y_k = A_i + B_j +$  residual model. Thus, the estimation procedure is simple:

- use the OLS to fit the additive model to the individual data
- create the variable  $e_k$  of the estimated residuals
- $\widehat{\text{Var}}(\hat{Y}_Q) = (N^2/n) \cdot (1/n) \sum_s e_k^2$ .

This formula is precisely that proposed in paragraph 2, and based on the super-population model. A rather neat situation!

#### 4.4.3 Discussion of the Results

The variance breaks down into two parts: one that can be seen as the probability of the square of the conditional bias; and one as the probability of the conditional variance.

The first term does not depend upon the quotas imposed on the sample, but only upon the quality of the fit of an additive model to the variable of interest. This part of the variance is diminished by choosing quota criteria that can best explain what we want to measure.

The second term, on the other hand, depends upon the remaining variability  $(N_{ij}^2 V_{ij}/n_{ij})$  and the number of observations collected in each cell. Since the size of the sample is fixed, we must attempt to make the  $n_{ij}$  as close as possible to Neyman's distribution:  $n_{ij} \propto N_{ij} V_{ij}^{1/2}$ . This may be achieved approximately by overloading quotas  $n_{i+}$  and  $n_{+j}$ , which correspond to large values of  $V_{ij}$ . Thus, in some cases, it is possible to improve the precision of a quota survey considerably.

## 4.5 Combination of the Quota Method and Stratified or Multi-Stage Samplings

### 4.5.1 The Case of Stratified Sampling with a Quota in Each Stratum

If the size of the criteria used to set the quotas are known in each stratum, the method described above makes it possible to construct an unbiased estimator, under the hypothesis that sampling functions like an SRS constraint in **each stratum**. If the allocation of quotas is proportional to the size of each stratum, the estimator is the natural estimator of the stratified sampling. If “national” quotas are used with each stratum, a correction should be made by reweighting.

On the other hand, if the size of the quota variables is unknown at the stratum level, it is not possible to correct the estimators to eliminate “structure effects” related to the stratification. Since, furthermore, the purpose of stratification is to construct dissimilar sub-populations, the corrections required will generally be quite large. Thus, the quota method is not recommended (except when the validity of the additive model is quite clear, *cf* part 3).

### 4.5.2 The Case of Two-Stage Sampling

Let us assume a two-stage sampling (inside a stratum where the sizes of the quota variables are known). If the sizes of the quota variables are known at the level of each primary unit, there are no problems. The theory in section 4.4 makes it possible to obtain an estimator of the total  $Y$  in each primary unit, as well as to calculate its variance, and an estimator of the latter. These quantities can then be used to obtain an estimator of  $Y$ , as well as an estimator of precision (*cf* Rao 1975). If the sizes of the quota criteria are not known at the level of the primary units, but only at the stratum level, we again have a problem that is impossible to correct. However, there is generally little harm if the PU are relatively similar: the structure of each PU is close to that of the stratum as a whole, and the corrections to be made for each PU are close to those that must be made at the stratum level.

### 4.5.3 In Conclusion

In conclusion, in the case complex multi-stage stratified sampling, the quota method may be used as the final sampling method if the stratification was carried out effectively by regrouping the similar primary units together, and if quotas derived from the data relative to each stratum are used with each PU.

To the extent that the hypothesis of simple random sampling constrained in each PU may appear to be quite satisfactory, the quota method is justified independently of any super-population model.

## 5. CONCLUSIONS AND PROBLEMS

### 5.1 How Should Non-response Be Taken into Account?

As we have already shown, this is the most important limitation in our theory. As far as sampling using the quota method is concerned, we do not have, in principle, any information on members of the population who refuse to respond to the survey, and we find ourselves lacking individual information on the subject of non-respondents. However, the situation is not as desperate as one might think. Let us illustrate this using a very simplified example.

We have carried out a simple quota survey using a sample of  $n_i$  individuals in category  $i$  with a population  $N_i$ . An acceptable model of non-response postulates a response probability of  $r_c$  if an individual belongs to category  $c$  with a population  $N_c$ . The (unknown) population



of the intersection between quota category  $i$  and class  $c$  of the non-response model is expressed as  $N_i^c$ . The population likely to respond in category  $i$  is thus  $N_{ri} = \sum_c N_i^c r_c$ . By setting a quota  $n_i$  in this category, within the framework of model (4.1), we obtain a probability of inclusion in the sample of  $w_i^{-1} = n_i/N_{ri}$ . In the sample, we collect  $n_i^c$  individuals belonging to the intersection  $(i, c)$  between the two categories. This quantity is random, and its probability is  $N_i^c r_c w_i^{-1}$ . If we attempt to estimate  $N_i^c$ , we will solve the estimating equations derived from the following relations:

$$N_i^c = n_i^c w_i r_c^{-1},$$

$$\sum_c N_i^c = N_i,$$

$$\sum_i N_i^c = N^c.$$

Thus, ranking ratio technique makes it possible to obtain estimates of  $\hat{r}_c$  and  $\hat{w}_i$ , and to derive estimators  $\hat{N}_i^c = n_i^c \hat{w}_i \hat{r}_c^{-1}$  from the sizes of the intersection  $(i, c)$ . We can also obtain an estimator of the total of  $Y$ :

$$\hat{Y}_{NR} = \sum_{ic} N_i^c \bar{y}_i^c = \sum_{ic} r_c^{-1} w_i n_i^c \bar{y}_i^c,$$

where  $\bar{y}_i^c$  is the mean of the  $Y_k$  values in the sample in category  $(i, c)$ . Thus, estimation techniques based on fitting should allow for the honourable processing of non-responses in quota surveys.

## 5.2 Some Points of Comparison with Probabilistic Surveys

Regardless of how we try to understand it, the quota method demands the formulation of a hypothetical model to fit the data. On the other hand, a probabilistic survey does not, in principle, depend upon any model. In practice, sampling for a probabilistic survey is a model to which the reality of data collection attempts to conform. In fact, we are well aware that, in any probabilistic survey, some compromises of detail must be made with the model (necessary exclusion of certain units, replacement of others after selection but before data collection, *etc*). However, we can say that statistical biases are always much lower in probabilistic selection than when using the quota method. On the other hand, quotas make it possible to use, in the sampling stage, additional information that cannot be mobilized in a probabilistic selection process. As a result, the variance of a quota sampling is similar to that of a regression estimation, and is thus generally smaller than that resulting from a probabilistic survey associated with its estimate of standard inflated values. The choice is between bias due to the model associated with low variance, against lack of bias. Two types of conclusions can be drawn from this approach:

**5.2.1** Precision depends mostly upon the size of the sample. On the average, in the case of small samples, probabilistic sampling will produce the worst results; and the bias of a quota survey will be more tolerable than the lack of precision of a probabilistic survey. For large samples, on the other hand, the quota method will have a clear bias that is obviously incompatible with the confidence interval without bias of a probabilistic survey.

Where should the boundary between the two methods be set? It is hard for the theory to be specific. On the other hand, experience in the French institutes may lead to a solution to this question: most national quota surveys are carried out on samples of 1,000 to 2,000 individuals. On the other hand, no national probabilistic survey mobilizes less than 5,000 units. It would seem fair to say that a size of 2,500 to 3,000 surveys is a practical boundary between the two types of surveys.

### 5.2.2 Official Statistics or Marketing

In a survey, the use of any speculative model represents methodological risk-taking. This may be perfectly reasonable if the users are aware of it, and if they have ratified the speculations leading to the specification of the model. This is typically what happens, at least implicitly, in marketing surveys: an organization, company, administration, or association requests a sampling survey from a polling company. A contract marks the agreement between the two parties respecting the implementation of the survey, its price, the result delivery schedule, and **the methodology used**. In this methodology, models are used to formalize the sampling or behaviour of the population. Thus, from this point of view, the use of the quota method may be quite proper.

Official statisticians, on the other hand, are responsible for generating data that can be used by the entire society; and that can be used, in particular, in the arbitration of disputes between various groups, parties, and social classes. The use of statistical models, particularly econometric models that describe the behaviour of economic agents, may turn out to be very dangerous, partial, or affected by a questionable or disputed economic theory. Official statistics should not tolerate any uncontrollable bias in its products. It should carry out sample surveys using probabilistic methods.

There is no real opposition between quota survey techniques and those using controlled randomness, quite the opposite – they are complementary. As a proof of this, the statistics that are used to construct the quotas are themselves very often derived from large surveys carried out by the National Statistics Services. However, quota survey technicians find it hard to admit that these data are obtained using methods other than traditional, confirmed, and well-founded probabilistic techniques.

### ACKNOWLEDGEMENTS

I would like to express my sincere thanks to the referee and editor for their help in improving the quality of this paper.

## APPENDIX

### Demonstration of the Results of Section 4.4

#### 1. Notation and Results

In order to deal with the question in a general way, we will require certain convenient notations. We have  $Q$  qualitative variables whose modalities are indicated by using indices from 1 to  $I_q$  when  $q = 1$  to  $Q$ . A “cell” is denoted as  $c$ ; that is, a series of  $Q$  indices where the  $q^{\text{th}}$  could have a value of 1 to  $I_q$ ; and  $q_c$  is the value of the  $q^{\text{th}}$  index ( $q^{\text{th}}$  projection of  $c$ ); in a finite population  $U$ , of size  $N$ ,  $U_c$  is the population of individuals in cell  $c$ , when the size of the cell is  $N_c$ . The quantity  $N_i^{+q} = \sum_{q_c = i} N_c$  is the total of the  $Q$ -dimensional contingency table where the cells are represented by  $c$  for the  $i^{\text{th}}$  modality of the  $q^{\text{th}}$  variable. If we postulate that

$$\bar{Y}_c = \frac{1}{N_c} \sum_{k \in U_c} Y_k.$$

We will obtain the following results:

**Result 1:** Variable  $Y_k (k \in U)$  may be parametrized by the following numbers:  $A_{q_c}^q, E_c$  and  $R_k$  by:

$$\bar{Y}_k = \bar{Y}_c + R_k \quad \text{if } k \in U_c. \quad \text{We have} \quad \sum_{U_c} R_k = 0 \quad \text{for any } c.$$

$$\bar{Y}_c = \sum_{q=1}^Q A_{q_c}^q + E_c \quad \text{with} \quad A_{I_q}^q = 0 \quad \text{for } q = 2 \text{ to } Q \quad \text{and}$$

$$\sum_{q_c=i} N_c E_c = 0 \quad \text{for } q = i \text{ to } Q \quad \text{and } i = 1 \text{ to } I_q.$$

These numbers are obtained from the minimization of:

$$\sum_U \left( Y_k - \sum_{q=1}^Q A_{q_c(k)}^q \right)^2 = \sum_c N_c \left( \bar{Y}_c - \sum_{q=1}^Q A_{q_c}^q \right)^2.$$

Let us assume that we have a sample  $s$ . We will use  $n$  to denote all quantities in the sample that are similar to whatever we have already indicated in the population.

We assume that  $s$  was obtained on the basis of simple random sampling (with or without replacement) in accordance with an equal probability scheme constrained by the totals  $n_i^{+q}$  ( $q = 1$  to  $Q$ ,  $i = 1$  to  $I_q$ ), the quotas.

The purpose of this appendix is to demonstrate the following result:

**Result 2:** The variance of  $\sum_c \hat{N}_c E_c$  is approximately equal to  $1/n \sum_c N_c E_c^2$  when  $n$ , and  $N/n$  become arbitrarily large.

The following section will provide a more precise formulation for this result.

#### 2. Sampling Plan and Asymptotic Reduction

Let us consider the following two sampling models SR and AR:

SR: Bernouilli Sampling. Each of the units of  $N$  belong to  $s$  with a probability  $f$ , and the  $N$  drawings are independent.



AR: Each unit is drawn a number  $v_k$  of times;  $v_k$  follows Poisson's law with  $f$  parameters. The  $v_k$  are independent variables.

A simple random survey without replacement (SRSWOR) of fixed size  $n$  is an SR sampling if the total size of the sample is  $n$ .

A simple random survey with replacement (SRSWR) of fixed size  $n$  is an AR sampling when we have  $n$  observations; that is, when  $\sum_k v_k = n$ .

In the case of SR sampling, the law of the vector  $n_c$  is obtained as follows:

$$\Pr(\{n_c\}) = \prod_c \binom{N_c}{n_c} f^{n_c} (1 - f)^{N_c - n_c}.$$

In the case of AR sampling, we have:

$$\Pr(\{n_c\}) = \prod_c \frac{(N_c f)^{n_c}}{n_c!} \exp(-fN_c).$$

In both cases the variables  $n_c$  are independent.

In the case of SR sampling constrained by  $\sum n_c = n$ , the law of the  $n_c$  is hypergeometric:

$$\Pr(\{n_c\}) = \prod_c \binom{N_c}{n_c} \binom{N}{n}^{-1}.$$

In the case of the restricted AR sampling, the law is multinomial:

$$\Pr(\{n_c\}) = \prod_c p_c^{n_c/n}.$$

The sampling plan model retained by the quota method described in paragraph 3 corresponds to constraints on these two schemes; which is equivalent to constraints on the SR and AR plans.

If we assume that  $N$  tends toward infinity, that  $f$  tends towards 0, and that  $n^* = fN$  tends toward infinity, then in the two plans, the law of the  $u_c = n^{*-1/2} (n_c - fN_c) = n^{*-1/2} (p_c^* - p_c)$ , with  $p_c^* = n_c/n^*$ , tends toward a multidimensional normal law with independent  $u_c$ , with zero probability and variances equal to  $p_c$ .

### 3. Proportional Sampling

In this case, we have  $\hat{N}_c = N/n$ , so that the quantity for which we want to determine the variance is:

$$\frac{N}{n^{*1/2}} \sum_c u_c E_c,$$

where the vector of the  $u_c$  follows a centered normal law with a diagonal covariance matrix  $\Delta = \text{diag}(p_c)$ , constrained by the relationships expressed by the quotas:

$$\sum_{q_c=i} u_c = 0 \text{ for } q = 1 \text{ to } Q, \quad i = 1 \text{ for } I_q \text{ if } q = 1, \quad i = 1 \text{ for } I_q - 1 \text{ if } q = 2 \text{ for } Q.$$

If we let  $U$  represent the vector of the  $u_c$ , the relationships can be written as follows:

$$AU = 0,$$

with  $A$  matrix with  $l = \sum_q I_q - (Q - 1)$  rows and  $k = \Pi_q I_q$  columns, where 1 and 0 represent the constraints. This also expresses the fact that  $U$  varies in the kernel  $L$  of the operator defined by matrix  $A$ . The (asymptotic) law of  $U$  is thus that of a centered gaussian vector  $W$  with a matrix whose covariances equal  $\Delta$ , when  $AW = 0$ . Thus, it is a question of evaluating the variance of a scalar product  $U'E$ , where  $E$  is the vector of the  $E_c$ .

It is important to emphasize the following two points:

- The constraints upon the  $E_c$  given in result 1 can be expressed on the basis of matrix analysis by  $A\Delta E = 0$ . In other words,  $\Delta E$  is a vector of  $L = \text{Ker}A$ , or a vector of  $\text{Ker}(A\Delta)$ .
- Let  $P$  be the projection of  $\mathcal{R}^k$  on  $L$  orthogonal in the  $\Delta^{-1}$  metrics.  $P$  verifies the following relations:
  - $\forall x \in L, Px = x; \text{Im } P = L$
  - $Px = 0 \Leftrightarrow \forall x \in L, x' \Delta^{-1} y = 0; \text{Ker } P = \Delta(L^\perp),$

where  $L^\perp$  is the supplementary line orthogonal to  $L$  in the natural metrics.

The gaussian vectors  $PW$  and  $(1 - P)W$  vary in  $L$  and  $\Delta(L^\perp)$  respectively; and their sum is equal to  $W$ . Moreover, they are independent; in fact, their covariance matrix is  $E(PW)((1 - P)W)' = P\Delta(1 - P')$ . Thus,  $P'$  is the kernel projector  $L^\perp$  and can be represented as  $\Delta(L^\perp)^\perp$ . The image of the projector  $(1 - P')$  is thus  $L^\perp$ . That of  $\Delta(1 - P')$  is  $\Delta(L^\perp)$ ; that is, the kernel of  $P$ , *q.e.d.*

At this point, we have to evaluate the variance of  $\sum_c u_c E_c = U'E$ . Thus, in accordance with the previous statements, we can write  $W = U + V$ , when  $U$  and  $V$  are independent. The law of  $W$  conditional upon  $W \in L$  is none other than the law of  $W$  conditional upon  $V = 0$ .

Moreover, we have:

$$V'E = (\Delta^{-1} V)'(\Delta E).$$

Since  $\Delta E$  is in  $L$ , and  $V$  varies in  $\Delta(L^\perp)$ , the scalar product above is zero. From this, we can deduce that:

$$\text{Var}(U'E) = \text{Var}(W'E) = E'E\Delta E = \sum_c p_c E_c^2.$$

The asymptotic variance of is thus equal to  $N/n^* \sum_c n_c E_c$

$$\frac{N^2}{n} \sum_c p_c E_c^2 = \frac{N}{n} \sum_c N_c E_c^2.$$

#### 4. Sampling using "Non-Proportional" Quotas

Let us complete the preceding asymptotic reduction. Now, the vector  $\hat{p}^\circ$  of  $n_c/n^*$  is constrained by

$$A\hat{p}^\circ = Ap + n^{*-1/2} AV_0,$$

where  $Ap$  is the vector (1-dimensional) of the “proportional quotas”, and  $V_0$  is the only vector ( $k$ -dimensional) of  $\Delta(L^\perp)$ , so that  $A(p + n^{*-1/2} V_0)$ ; that is, the vector of the quotas imposed. Thus, as in the previous paragraph,  $U = n^{*1/2} (\hat{p}^\circ - p)$  may be analyzed as a gaussian vector  $W = U + V$  conditional upon  $V = V_0$ . Thus,  $EU_0 = V_0$ , and the covariances matrix of  $U_0$  is the same as that of  $U$ .

Moreover, we go from  $\hat{p}^\circ$  to  $\hat{p}$  by estimating the maximum resemblance. Under asymptotic gaussian conditions, this consists of minimizing the quadratic form  $(\hat{p}^\circ - \hat{p})' \Delta^{-1} (\hat{p}^\circ - \hat{p})$  under constraints  $A\hat{p} = Ap$ . Since  $\hat{p}^\circ$  varies in the related subspace  $L + V_0$  that is parallel to  $L$ , and minimization is a question of projecting  $\hat{p}^\circ$  upon  $L$  orthogonally for  $\Delta^{-1}$ ; that is, along  $\Delta(L^\perp)$ , it follows that we have  $\hat{p} = \hat{p}^\circ - n^{*-1/2} V_0$  under asymptotic conditions. The random vector  $\hat{p}$  is thus obtained from  $\hat{p}^\circ$ , is unbiased, and has the same covariance matrix as  $\hat{p}^\circ$ , so that  $n^{*-1/2} U$ .

Finally, we have:

$$E \left( \sum_c \hat{p}_c E_c \right)^2 = E(\hat{p}' \underline{E})^2 = \frac{1}{n^*} \sum_c p_c E_c^2$$

as in the previous case.

### REFERENCES

CASSEL, C.M., SÄRNDAL, C.E., and WRETMAN, J.H. (1977). *Foundations of Inference in Survey Sampling*. New York: Wiley & Sons.

COCHRAN, W.G. (1977). *Sampling Techniques*. New York: Wiley & Sons.

DESABIE, J. (1965). *Théorie et pratique des sondages*. Paris: Dunod.

DEVILLE, J.C., and SÄRNDAL, C.E. (1990). Calibration estimators and generalized raking techniques. Manuscript submitted for publication.

GOURIÉROUX, C. (1981). *Théorie des sondages*. Paris: Economica.

MADOW, W.G., OLKIN, I., and RUBIN, D.B., (Eds.) (1983). *Incomplete Data in Sample Surveys*. New York: Academic Press.

RAO, J.N.K. (1976). Unbiased variance estimation for multistage designs. *Sankhyā*, Series C, 37, 133-139.

SMITH, T.M.F. (1983). On the validity of inferences from non-random samples. *Journal of the Royal Statistical Society*, A, 146, 394-403.

WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.





# Sampling Flows of Mobile Human Populations

GRAHAM KALTON<sup>1</sup>

## ABSTRACT

Surveys are often conducted of flows of persons, such as: visitors to museums, libraries and parks; voters; shoppers; hospital outpatients; tourists; international travellers; and car occupants. The sample designs for such surveys usually involve sampling in time and space. Methods for sampling flows of human populations are reviewed and illustrated.

**KEY WORDS:** Mobile populations; Exit polls; Traffic surveys; Time and space sampling; Systematic sampling.

## 1. INTRODUCTION

Most surveys of human populations are household based, typically with a sample of households selected with a multi-stage sample design, and individuals sampled within the selected households. The household survey is a powerful method for collecting data on a wide range of characteristics about the population, such as social, demographic, economic and health characteristics and the population's opinions and attitudes. The method is, however, not so effective for studying the characteristics of mobile populations. Two types of mobile populations may be distinguished: those who do not reside regularly at a fixed location, such as nomads and the homeless; and members of the general population who belong to the mobile population under study because they are in transit, such as visitors to libraries and parks, voters at polling booths, shoppers, hospital outpatients, travellers, and car occupants. This paper reviews sample design issues for this latter type of mobile population.

Although there are many surveys concerned with flows of mobile human populations, the general sampling literature contains little discussion of the sampling issues involved. The purpose of this paper is to describe the sample designs commonly adopted for surveys of flows of human populations, to discuss some of the special sampling issues faced, and to illustrate the range of applications for such surveys. The next section of the paper reviews the general time and space sample design used for sampling persons in transit and some of the issues involved in employing this design in particular situations. Section 3 then illustrates the application of the design in a range of different settings. Section 4 presents some concluding remarks.

## 2. SAMPLING IN TIME AND SPACE

It will be useful to consider a specific example in describing the general time and space sample design for sampling flows of human populations. Suppose that a survey of visitors to a summer sculpture exhibition in a city park is to be conducted to find out the visitors' socio-economic characteristics, how they heard about the exhibition, what means of transport they used to get to the park, and perhaps their views of the exhibition. Suppose that the exhibition is held from April 1 to September 30 in the year in question, that it is open from 10 a.m. until 6 p.m. daily, and that there are three sites where visitors enter and leave the exhibition.

---

<sup>1</sup> Graham Kalton, Survey Research Center, University of Michigan, Ann Arbor, Michigan 48106-1248, U.S.A.

The sampling frame for a survey of this type is usually taken to be a list of time interval/site primary sampling units (PSUs). This frame is constructed by dividing the time period of the survey into a set of time intervals for each site. A simple construction of PSUs for the current example would be to divide each exhibition day at each site into two time intervals, one from 10 a.m. until 2 p.m. and the other from 2 p.m. until 6 p.m. A more complex construction of PSUs could involve time intervals of different lengths on different days and/or at different sites. Once the PSUs are defined, a two-stage sample design is often employed. At the first stage a sample of PSUs is selected, and at the second stage a sample of visitors is drawn, usually by systematic sampling, in the sampled PSUs.

The actual specification of the sample design for a survey of persons in transit within the two-stage sampling framework depends on features of the mobile population under study and of the survey data collection procedures. A key feature is the nature of the flow of the mobile population. In particular, is there a predictable variability in the rate of flow across PSUs? For instance, is the flow at one site higher than that at another site, or are the flows at some time intervals (say, Saturday afternoons) higher than those at others? Also, is the flow within a PSU a smooth one throughout the time interval or is it uneven, with visitors arriving (or leaving) in sizeable groups? Both these aspects of flow affect the sample design for the survey.

If the flow is fairly uniform across the PSUs, and if the PSU time intervals are the same, then the number of visitors per PSU is approximately constant. In this case, the PSUs may be sampled with equal probabilities, and a constant subsampling fraction can be applied within the selected PSUs to generate an equal probability, or *epsem*, sample of visits. The PSUs can be classified in two or more dimensions (*e.g.* day of week, time of day, and site), and a carefully balanced sample across these dimensions can be obtained using lattice sampling (Yates 1981; Cochran 1977 and Jessen 1978).

In many cases, the level of flow varies across the PSUs in a manner that is partly predictable. For instance, the attendance at the sculpture exhibition may be known to be generally higher in the later shift each day and at the weekends, and particularly low on Mondays. Thus the PSUs comprise different numbers of visitors, that is, they are PSUs of unequal sizes. The usual procedure for handling PSUs of unequal sizes is to sample them with probabilities proportional to their sizes (PPS), or estimated sizes (PPES). In the current context, the actual PSU sizes are not known in advance, and estimated sizes must therefore be used. Sampling the PSUs with PPES works well provided that reasonable estimates of the sizes can be made. When PSUs are selected by PPES sampling, then the application within the selected PSUs of subsampling fractions that are inversely proportional to the estimated sizes of the PSUs produces an overall *epsem* sample of visits. In general, an attraction of PPES sampling (with reasonable estimates of size) is that the subsample sizes in the PSUs do not vary greatly from one PSU to another. This feature is of especial value for conducting the fieldwork in surveys of persons in transit. When time/site PSUs are sampled by PPES sampling, lattice sampling cannot be applied for deep stratification. Instead, controlled selection may be employed for this purpose (Goodman and Kish 1950; Hess *et al.* 1975).

An important consideration in any two-stage sample design is the allocation of the sample between first-and second-stage units, that is, how many PSUs to select and how many elements to select per sampled PSU. In the case of surveys of persons in transit, that allocation is strongly affected by the fieldwork procedures to be used and the nature of the flow within the PSUs. The aim of the design is to make full use of the fieldworkers assigned to a sampled PSU while maintaining a probability sample of persons entering (or leaving) the site during the sampled time interval.



Many surveys of persons in transit use self-completion questionnaires, in which case the fieldwork process for the two-stage design described above consists of counting persons as they enter (or leave) the sampled site during the time interval, selecting every  $k$ th person for a systematic sample, and asking the selected persons to complete the questionnaire. If the flow is light and evenly spread throughout the time interval, one fieldworker may be able to handle all the tasks involved. When this is so, the sampling interval  $k$  can be chosen to give the fieldworker time to perform all the tasks in an unpressured way. If, however, the flow is heavy, either constantly or intermittently, two fieldworkers may be needed, one simply to count entrants (or leavers) and identify sampled persons, and the second to hand out the questionnaires and to instruct respondents on how they should be completed and returned. With this fieldwork arrangement, the sampling interval can be chosen to keep the second fieldworker as fully occupied as possible, while making sure that he or she is able to distribute questionnaires to all (or at least nearly all) of those sampled. Nonresponse can be a major concern with the self-completion mode of data collection. It is often possible to keep nonresponse to an acceptable level when sampled persons complete and return the questionnaire at the site. However, when they are handed the questionnaire with the request to complete it later and return it by mail, the level of nonresponse can be very high and, moreover, there is generally no way of following up the nonrespondents.

When face-to-face interviewing is used for data collection, the fieldwork team for a PSU usually contains one counter and a small team of interviewers. The size of the interviewer team depends on the regularity of the flow and the length of the interview. Since persons in transit are likely to be unwilling to be delayed for long, interviews are necessarily mostly short. Longer interviews may, however, be possible if the sampled persons are in a waiting mode, such as waiting in line or in an airport departure lounge. The choice of sampling interval has to be such that there is always (or nearly always) an interviewer free to interview the next sampled person, and that the interviewers do not spend too much time waiting for the next sampled person to be selected. If the flow is irregular, allowance needs to be made to accommodate the peaks (for instance, the arrival of a coachload of visitors to the sculpture exhibition).

The PPES selection of the PSUs works to equate the subsample size for each sampled PSU. For face-to-face interview surveys, the interviewer load is thus roughly the same for each selected PSU, and hence the same-sized interviewer team can be used for each PSU. A problem occurs, however, when the PPES measure used in selecting the PSU at the first stage is seriously in error. For example, a thunderstorm may substantially reduce the number of visitors to the sculpture exhibition on a particular Saturday afternoon, or an unforeseen holiday may substantially increase the number on another day. In the first case, applying in that PSU a sampling interval inversely proportional to its estimated size will leave the interviewers largely unoccupied, whereas in the second case it will result in a workload that the interviewers cannot handle. A modification that may be adopted in such cases is to change the sampling interval at the start of data collection to one that is more suitable for the flow actually encountered. Since this modification destroys the *epsem* property of the sample, weights are needed in the survey analysis.

A general limitation to the systematic sampling of visitors at selected PSUs is that if the sampling interval is made long enough to enable interviewers to cope with peak flows, they spend much of their time without work. On the other hand, if the sampling interval is reduced, the interviewers are more fully occupied, but they cannot cope with peak flows. Various methods have been proposed to circumvent these problems (Heady 1985). One procedure is to take a systematic sample of times (say, every 10 minutes) and to select the next visitor to enter after each sampled time. This procedure might have fieldwork attractions, but it does

not produce a probability sample of visitors. Persons arriving in busy periods are less likely to be chosen, as are those who travel in groups, and the walking habits of persons travelling in groups may affect the chances of selection in unknown ways. The sample generated by this procedure is clearly not an epsem sample. An attempt can be made to compensate for the selection bias that operates against visitors arriving in busy times by dividing the time interval for selected PSUs into a set of much shorter intervals, and keeping a log of arrivals in each such interval. Then weighting adjustments can be employed to compensate for the variation in the flow across the shorter intervals.

Another alternative procedure to systematic sampling of visitors is to take the next person to enter (or leave) after the last interview was completed. With this procedure, the first persons to arrive after gaps in the flow, perhaps the leaders of groups, clearly have greater chances of selection. Also interviewers may deliberately speed up or slow down their current interview in order to avoid or to select a particular individual. For these reasons, variants on this procedure that select the  $n$ th person after the completed interview, where  $n$  might be set at 2, 3, 4 or 5, have been employed. These alternatives to straightforward systematic sampling of visitors make more effective use of interviewers' time, and hence enable larger samples to be obtained for a given fieldwork budget. However, they produce nonprobability samples, with the risk of selection bias that this form of sampling entails. Probability sampling provides the security of objective statistical inference without the need for assumptions about the sample selection process. With nonprobability sampling, assumptions need to be made about the way the sample was generated, a common assumption being that all the elements in the population have an equal chance of selection. Failure of the assumptions can lead to serious bias in the survey estimates.

Visitors may be sampled either as they enter or as they leave a location. If data about the visitors' activities in and opinions of the location are required, then leavers need to be sampled. In other cases, the choice between sampling entrants and leavers may depend on the nature of the flows. It may, for example, be difficult to sample and interview people leaving a theatre because they leave en masse and because they will not want to be delayed. On the other hand, they may be readily sampled and interviewed as they line up to enter the theatre.

In concluding this section, attention should be drawn to the fact that the samples described here are samples of visits not visitors. The standard two-stage design may produce an epsem sample of visits, but this is not the same as an epsem sample of visitors unless each visitor visits the place under study (the sculpture exhibition) only once (or they all visit the same number of times). For most flow surveys, the visit, rather than the visitor, is the appropriate unit of analysis. There are, however, situations where the analytic unit is problematic. Using the visit as the unit of analysis, the researcher might readily accept visits to the sculpture exhibition on two separate days as distinct visits, but might not be willing to treat two entries on the same day (one, perhaps, after leaving briefly for refreshments) as two visits. The use of the visitor as the unit of analysis presents severe problems because of the issue of multiple visits, and the fact that visitors will not be able to report their multiplicities. They may be able to recall past visits reasonably well, but they will usually be unable to forecast future visits accurately.

### 3. SOME EXAMPLES

This section presents some examples of surveys of flows of human populations in order to indicate the wide range of applications and to illustrate some of the special considerations that arise in particular settings.



### 3.1 A Survey of Library Use

A survey of the use of the 18 libraries at the University of Michigan was conducted in 1984 (Heeringa 1985). Each sampled person exiting a library was asked whether he or she had used the library's materials and services during that visit. If so, the person was asked to complete a short self-completion questionnaire of seven questions on the materials and services used. Most of the 5,184 respondents completed the questionnaires on the spot and returned them to the survey fieldworkers; others sent them back by campus mail. A response rate of 96% was obtained.

The sample design followed the two-stage time/site sample design described in Section 2. The survey covered the full 1984 calendar year. Each day the libraries were open was divided into 10 two-hour time intervals, starting at 7.30 a.m. and lasting until 3.30 a.m. the next morning, the two-hour interval being chosen on the grounds that it was a suitable shift for the fieldworkers. The PSUs were then defined to be time interval/library combinations. The PSUs were selected by PPES sampling, where the estimated size for a PSU was the estimated number of persons exiting from that library in the specified time period. Rough estimates of these numbers were derived from average daily usage based on November, 1983, turnstile counts where available, and on librarians' estimates where not, and on an assumption that library exit volume was twice as high between 9.30 a.m. and 5.30 p.m. as at other times. The libraries were stratified into four types, and within each stratum controlled selection was employed to give a proportionate distribution of the sample across libraries, days of the week, and time intervals.

For each selected PSU, a systematic sample of persons exiting the library was selected for the survey, with the sampling interval being determined to yield an overall epiem sample of visits. Fieldworkers were provided with a record sheet of integers from 1 up to 430, with the selected numbers marked on them. All they then needed to do was check off a number for each person exiting the library, and select the persons associated with the sample numbers. An advantage of this scheme is that fractional sampling intervals are readily handled. Where the exit volume for a sampled PSU was expected to be low, one fieldworker was assigned to perform both the counting and the contacting of sampled persons. Where the exit volume was high, two fieldworkers were assigned, one to count and one to contact sampled persons. There was also a need for more than one fieldworker for libraries with more than one exit.

### 3.2 A Survey of Museum Visits

A face-to-face interview survey of visitors leaving the National Air and Space Museum in Washington, D.C. was conducted from mid-July until December, 1988 (Doering and Black 1989). The interview, which took about four to six minutes to complete, collected data on the sampled person's socio-demographic background, place of residence, activities on the visit, exhibits of special interest, reason for visit, the size and type of group if part of a group visit, and mode of transport used. Children under 12 years old and persons working at the museum were excluded from the survey. Data were collected from 5,574 respondents, with a response rate of 86%.

Each day in the survey period was divided into two half-days. Interviewing was conducted on one half-day every second day, alternating between mornings and afternoons. During the summer season, three public exits from the museum were in operation, while later in the year only two of them were open. During the selected half-days, survey data collection was rotated on an hourly basis between the exits that were open.



The fieldwork team for an exit at a sampled hour comprised one or two counters and two interviewers. The lead counter used a mechanical counter and a stop watch to keep track of the number of persons exiting, and to maintain a record that gave the numbers of persons exiting in each 10-minute interval in the hour. The lead counter also identified the persons to be interviewed. The selection of sample persons was made in order to keep the interviewers fully occupied. The lead counter noted when an interviewer had completed an interview and was ready to begin another one, and then chose the fifth person exiting after that time as the next sampled person. The 10-minute flow counts were used in the analysis to develop weights to compensate for the variation in the chance of selection associated with the variable flow of persons across time.

The distinction between the “visit” and the “visitor” is particularly salient for this survey. Persons could, of course, visit the museum on several days throughout the survey period, and also could visit the museum several times on a given day. This latter possibility is particularly likely with the National Air and Space Museum because entry to the Museum is free, and hence there is no incentive to enter only once. Given this situation, it may be appropriate to define multiple entries on one day as a single visit for some types of analysis. For some purposes, this definition could be applied by restricting the analysis to those exiting for the first time on the sampled day.

### 3.3 Exit Polls

A number of major news organizations conduct polls of voters on election days in the United States (Levy 1983; Mitofsky 1991). Voters are sampled as they leave polling places. Those selected are asked to complete a short and simple self-completion questionnaire, and to deposit the completed questionnaire in a ballot box. A typical questionnaire contains around 25 questions asking how the respondent voted, what the respondent’s position is on key issues, what opinions the respondent has on various topics, and what are the respondent’s demographic characteristics. Refusal rates for the CBS exit polls have averaged 25% for recent elections (Mitofsky and Waksberg 1989).

The sampling of voters for election polls usually employs a straightforward two-stage sample design. At the first stage a stratified PPES sample of voting precincts is drawn, where the size measure is the number of voters in the precinct. At the second stage a systematic sample of voters leaving the polling place is selected, with a sampling interval chosen to produce an approximately epiem sample of voters within states. Usually only one interviewer is assigned to each selected precinct. The fieldwork is straightforward when a polling place has a single exit, and the interviewer is permitted to get close to it. When there are two or more exits, interviewers alternate between the exits, covering each one for set periods of time. When this applies, the sampling interval has to be modified accordingly. In some states interviewers are not allowed to approach within a certain distance of a polling place, and this can create problems if it results in voters departing in different directions before the interviewer can contact them.

### 3.4 Ambulatory Medical Care Survey

The U.S. National Ambulatory Medical Care Survey (NAMCS) employs a flow survey design to collect data on visits to physicians’ offices for physicians in office practice who direct patient care (Bryant and Shimizu 1988). The NAMCS has been conducted a number of times since it was introduced in 1973. For each survey, data collection has been spread throughout the survey’s calendar year in order to provide annual estimates of visit characteristics. Individual sampled physicians have, however, been asked to provide information for a sample of their visits occurring in only one week. The annual coverage is achieved by asking different sampled physicians to report on different weeks of the year.

The sample for the NAMCS is based on a complex three-stage design, which has varied over time. A broad overview of the design will serve for present purposes; for more details, the reader is referred to Bryant and Shimizu (1988). The first stage of the NAMCS sample design is the selection of a stratified PPES sample of areal PSUs, selected with probability proportional to population size. At the second stage, physicians are sampled from lists within the selected PSUs with different sampling intervals from PSU to PSU to take account of the unequal selection probabilities for the PSUs (in the more recent surveys, different specialty classes are sampled at different rates). Sampled physicians are then assigned at random in a balanced way to one of the 52 reporting weeks of the year. Each physician is asked to record information for a systematic sample of his or her patient visits occurring during the sampled week, with the sampling interval being chosen to yield about 30 sampled visits in the week. A sampling interval of 1, 2, 3 or 5 is chosen for a particular physician on the basis of the number of office visits the physician expects during the week, and the number of days he or she expects to see patients. The fieldwork procedures consist of keeping a log of patient arrivals for sampling purposes, and then completing a short 16-item record for each sampled visit.

The NAMCS is a survey of patient visits not patients. As such, it provides useful information about the nature of physicians' work on a visit basis – the frequency of use of diagnostic tests, the therapies provided, and the demographic characteristics of the patients seen. It does not, however, provide estimates on a patient basis, such as treatments and outcomes for patients' episodes of illness.

### 3.5 Surveys of International Passengers

A number of countries conduct surveys of their international travellers, both those entering and those leaving the country by land, sea or air. This subsection will briefly describe the sample designs for a survey of international air passengers conducted by the United States, for surveys of international air and land travellers conducted by Canada, and for a survey of international air and sea passengers conducted by the U.K.

The United States Travel and Tourism Administration conducts an In-flight Survey of International Air Travelers to survey both foreign travellers to the U.S. and U.S. residents travelling abroad (see, for instance, United States Travel and Tourism Administration 1989). The survey is conducted through the voluntary cooperation of some thirty airlines. A stratified sample of scheduled flights is selected for the third week of each month and all passengers on those flights are included in the sample. Participating airlines are provided with a survey kit of instructions and questionnaires in appropriate languages for each sampled flight. The airline cabin personnel distribute the self-completion questionnaires in boarding areas or in flight to all adult passengers and collect them prior to debarkation. Nonresponse is a serious problem with these surveys. For the 1988 survey of visitors to the United States, one half of the flight kits issued resulted in no returned questionnaires. For flights for which questionnaires were returned, the estimated response rate for non-U.S. residents was 44% and for U.S. residents it was only 20%.

The International Travel Section of Statistics Canada conducts international travel surveys at both airports and landports in Canada. The surveys are undertaken in cooperation with Canada Customs, with customs officers being responsible for distributing the self-completion mail-back questionnaires. The account here is based on the report by the International Travel Section, Statistics Canada (1979). It reflects the survey designs that applied prior to some changes that have recently been made. The sample designs for the landports and airports have been similar, and therefore only the design for the landports will be outlined here.

At one time the sampling scheme at landports for returning Canadian residents who had spent at least one night abroad was to distribute survey questionnaires to every travel party



on every fourth day throughout the year, the days being chosen by systematic sampling. This scheme proved to be unworkable because the customs officers too often failed to apply it correctly. It was therefore replaced by a stint scheme in which a landport was assigned two periods, or stints, for each quarter of the year during which the questionnaires were to be distributed. The stints were expected to last from 6 to 10 days, with successive stints starting about 6½ weeks apart (Gough and Ghangurde 1977). The number of questionnaires sent to a landport for a particular stint was determined from the expected traffic at that port. The customs officers were then instructed to start the distribution of the questionnaires on a given day, and to continue to distribute them until none were left. This sample design is geared to operational limitations resulting from the use of customs officers, for whom the survey is of only secondary concern, as survey fieldworkers. The design has some major drawbacks, but perhaps a more serious concern is a response rate of 20% or less.

The U.S. and Canadian surveys of international travellers both rely on cooperation from other agencies in conducting the fieldwork. This cooperation has notable benefits in costs, but a price is paid in terms of a lack of ability to apply rigorous controls to the fieldwork procedures. The U.K. surveys of air and sea travellers employ more costly face-to-face interviewing procedures.

The 1984 U.K. International Passenger Survey included the three Heathrow terminals, Gatwick and Manchester airports as strata (Griffiths and Elliot 1987). Within each airport, days were divided into mornings and afternoons, and these periods constituted the PSUs. A stratified sample of PSUs was selected, and systematic samples of passengers were chosen in selected PSUs. A sample of PSUs for other airports was also included. Two alternative data collection procedures were used at seaports. At some seaports, interviewers sampled and interviewed passengers at the quayside. At others, the interviewers travelled on the ship, interviewing passengers during the voyage. In the former case, they worked shifts that covered several sailings, and the shift became the PSU. In the latter case, the crossings were the PSUs.

### 3.6 Surveys at Shopping Centers

Surveys conducted at shopping centers are of two types. One type aims to describe the shoppers' socio-economic characteristics, their areas of residence, and their shopping activities in the center. The other type uses the shopping center as a convenient location to obtain samples of people from the general population of the area.

An example of a survey of the first type is a study that was conducted to examine the impact of the opening of a hypermarket on the outskirts of the city of Southampton, England (Wood 1978). Surveys of shoppers were conducted in four neighboring shopping centers both before and after the hypermarket opened (and also at the hypermarket). At each center, the first step in the survey process was the enumeration of all the retail outlets and their hours of opening. The second step was a counting of departures of groups of shoppers from sampled shops at sampled hours, with counting being conducted for 15 minutes within the hour. The counting operation was carried out over a period of one month. Based on the counts obtained, interviews were allocated between shop types and days of the week, and to specific shops and hours. Interviewers were then instructed to interview the given number of people leaving the shop, interviewing the next person to leave after they had completed the previous interview. The sample is one of shop visits, and shoppers could visit several shops on a particular trip to the shopping center. Respondents were asked about previous visits to shops in the center on this particular trip, and also about the number of extra shops they planned to visit. These data were used to develop weights for analyses of trips.



The second type of shopping center survey uses the selected persons at shopping centers as a convenience sample of the general population. Mall intercept surveys of this type are widely used in market research (Bush and Hair 1985; Gates and Solomon 1982). The procedures are often haphazard, and the samples are potentially biased. The issues involved are reviewed by Sudman (1980), who discusses procedures for sampling shopping centers, locations at selected centers, and time periods to improve the sample designs, and by Blair (1983), Dupont (1987), and Murry *et al.* (1989).

### 3.7 Road Traffic Surveys

One form of road traffic survey relates to traffic passing through one or more locations. Time and space sample designs can be applied for these surveys in a relatively straightforward manner. Kish *et al.* (1961), for example, describe the sample design for an origin-destination survey of vehicles using the Port of New York Authority's bridges over and tunnels under the Hudson river during 1959. A four-stage stratified PPES sample design was used for this survey. The PSUs were combinations of eight-hour shifts and particular bridges or tunnels. A sample of these PSUs was selected at the first stage, a sample of contiguous toll lanes (locations) was selected at the second stage within selected PSUs, a sample of specific lanes was selected at the third stage within selected locations, and finally a systematic sample of vehicles was selected at selected lanes. Interviewers stayed at one sampled location for four hours, and moved each hour from one traffic lane to another according to a prescribed pattern.

Another type of road traffic survey relates to general traffic on the road. Surveys of occupants of passenger vehicles to study seat belt usage and drivers' blood alcohol concentrations are of this type. A full discussion of the complex design issues involved in such surveys is outside the present scope; instead only a few general observations will be made.

The method of data collection to be employed exerts a strong influence on the sampling procedures for a general traffic survey. Seat belt usage is mostly studied by observational methods, whereas the measurement of blood alcohol concentrations usually involves breathtesting. Shoulder belt usage of front-seat occupants can be observed in moving traffic, but lap belt usage and the seat belt usage of other occupants can be observed only when the vehicle has stopped briefly, for instance at traffic lights. Lack of street lights can preclude observation of seat belt usage at night at some sites. Breathtesting requires the vehicle to be stopped, and this can be done safely only in locations where the stopped vehicle does not hinder the other traffic. Unlike observational surveys, interview surveys that stop vehicles face a significant nonresponse problem.

An ingenious method of studying seat belt usage on interstate highways is described by Wells *et al.* (1990). For this study, an observer sat behind the driver in a passenger van that travelled at a slower speed than the prevailing traffic in the right hand lane of the highway. From that vantage point, the observer noted the shoulder belt usage of front-seat occupants of cars, light trucks, and vans that passed the observer's van in the adjacent lane.

A more usual approach to studying seat belt usage is to take observations at road intersections and freeway exits controlled by traffic lights, and sometimes at shopping centers and parking lots (Ziegler 1983; Bowman and Rounds 1989). O'Day and Wolfe (1984) describe an observational survey of seat belt use in Michigan applying this approach. They sampled a number of areal units, sampled a number of intersections with traffic signals within these areas, sampled days for observations to be taken at these intersections, and sampled five periods of one hour each between 8 a.m. and 8 p.m. for observation on each selected day. Each hour of observation was conducted at a different intersection. The hours were selected by a scheme that alternated one hour working and one hour free, with the observers moving between

intersections in the free hours. Observations of seat belt usage were taken at the selected intersections at the specified times for vehicles that stopped at the traffic lights. When more than one vehicle was stopped, observation began with the second vehicle, because of the bias associated with the first vehicle to stop at a light. In order to obtain more detailed information on the usage of child-restraints, observations were also made on vehicles entering shopping centers and rest areas.

The usual approach to analyzing observational data on seat belt usage is to calculate the proportion wearing seat belts among those observed. Brick and Lago (1988) propose an alternative measure, the proportion of estimated time front-seat occupants are belted in eligible vehicles to the total time in eligible vehicles. For their survey a probability sample of all roadway intersections, whether they had traffic signals or not, was selected. To avoid selection bias, observers were told the site they were to use for observation and the direction of the traffic to be observed in the specified 40-minute interval of observation. The time occupants were on the road was estimated as the length of the road segment leading to the intersection divided by the estimated average speed of the traffic on that segment. This estimated time was used as a weighting factor in the analysis.

The sampling considerations for roadside breathtesting surveys are broadly similar to those for seat belt usage surveys, except that the locations for data collection need to be places where vehicles can be stopped safely. In the 1986 U.S. National Roadside Breathtesting Survey, local police officers cooperated in the survey by flagging down selected drivers and directing them to the survey interviewers (Wolfe 1986). The interviews lasted about 5-6 minutes. When an interviewer finished an interview and the respondent had taken the breath test, the interviewer would signal to the police officer to stop the next passing vehicle. Interviewing was conducted for a period of two hours at each sampled location. A count was made of all the vehicles passing the location in the sampled direction during the period, and the ratio of this count to the number of interviews conducted was used as a weighting factor in the analysis.

#### 4. CONCLUDING REMARKS

As the examples in the previous section illustrate, fieldwork considerations and the economics of data collection play major roles in the choice of sample design for surveys of persons in transit. The length of the time interval used in defining the PSUs may, for instance, be dictated by the length of a suitable workshift for the fieldworkers, and this may result in PSUs with substantial internal variation in the rate of flow. For example, in a survey of passengers arriving at a railway station, a morning interviewer workshift may include a peak flow of early morning commuters and a low rate of flow later on. If it were not for the need to make the PSU time interval conform to the fieldworkers' workshift, it would be preferable to avoid such variation in flow within PSUs since it leads to problems in how to subsample in the selected PSUs.

When the flow of persons within a PSU is uneven, the use of systematic sampling, or any epsem sampling scheme, for selecting persons creates a variable workload over time. If this variability in workload is substantial, there are difficulties in deciding how to staff the PSU for the survey fieldwork, particularly for a face-to-face interview survey. The assignment of sufficient staff to cope with peak flows is uneconomic since interviewers will then often be inactive at off-peak times. Sometimes staffing for somewhat below peak flow may be preferable. This will introduce some nonresponse at times of peak flow because no interviewer is available to conduct an interview with some sampled persons, but it will more fully use the interviewers' time.



The most effective use of the interviewers' time is to assign them to interview the first person to arrive (or leave) after they have completed their current interview. Schemes of this type suffer the disadvantage of not producing probability samples, and hence there is a risk of bias in the survey estimates. Where cost effective probability sampling designs can be devised, they are to be preferred. However, the choice of a sampling scheme in which the first (or second, or third) person is selected after an interviewer becomes free is understandably attractive for face-to-face interview surveys when the flow is very variable and unpredictable. When this kind of scheme is employed, it is useful to take counts of the flow over short intervals of time. These counts may then be used to make weighting adjustments to compensate for the unequal selection probabilities caused by the uneven flow.

### ACKNOWLEDGEMENT

I would like to record my thanks to many researchers who generously provided me with information about the flow surveys with which they have been associated.

### REFERENCES

- BLAIR, E. (1983). Sampling issues in trade area maps drawn from shopper surveys. *Journal of Marketing*, 47, 98-106.
- BOWMAN, B.L., and ROUNDS, D.A. (1989). Restraint System Usage in the Traffic Population, 1988 Annual Report. Washington, D.C.: U.S. Department of Transportation.
- BRICK, M., and LAGO, J. (1988). The design and implementation of an observational safety belt use survey. *Journal of Safety Research*, 19, 87-98.
- BRYANT, E., and SHIMIZU, I. (1988). *Sample Design, Sampling Variance, and Estimation Procedures for the National Ambulatory Medical Care Survey*. Vital and Health Statistics, Series 2, No. 108, Washington D.C.: U.S. Government Printing Office.
- BUSH, A.J., and HAIR, J.F. (1985). An assessment of the mall intercept as a data collection method. *Journal of Marketing Research*, 22, 158-67.
- COCHRAN, W.G. (1977). *Sampling Techniques* (3rd ed.). New York: John Wiley.
- DOERING, Z.D., and BLACK, K.J. (1989). Visits to the National Air and Space Museum (NASM): Demographic Characteristics. Working Paper 89-1, Institutional Studies, Smithsonian Institution.
- DUPONT, T.D. (1987). Do frequent mall shoppers distort mall-intercept survey results? *Journal of Advertising Research*, 45-51.
- GATES, R., and SOLOMON, P.J. (1982). Research using the mall intercept: State of the art. *Journal of Advertising Research*, 22, 4, 43-49.
- GOODMAN, R., and KISH, L. (1950). Controlled selection - a technique in probability sampling. *Journal of the American Statistical Association*, 45, 350-372.
- GOUGH, J.H., and GHANGURDE, P.D. (1977). Survey of Canadian residents returning by land. *Survey Methodology*, 3, 215-231.
- GRIFFITHS, D., and ELLIOT, D. (1987). Sampling errors on the International Passenger Survey. Unpublished paper, Social Survey Division, U.K. Office of Population Censuses and Surveys, London.
- HEADY, P. (1985). Note on some sampling methods for visitor surveys. *Survey Methodology Bulletin*. U.K. Office of Population Censuses and Surveys, 17, 10-17.
- HEERINGA, S.G. (1985). The University of Michigan 1984 Library Cost Study: Final Report. Institute for Social Research, University of Michigan.



- HESS, I., RIEDEL, D.C., and FITZPATRICK, T.B. (1975). *Probability Sampling of Hospitals and Patients*. Ann Arbor, Michigan: Health Administration Press.
- JESSEN, R.J. (1978). *Statistical Survey Techniques*. New York: John Wiley.
- KISH, L., LOVEJOY, W., and RACKOW, P. (1961). A multi-stage probability sample for traffic surveys. *Proceedings of the Social Statistics Section, American Statistical Association*, 227-230.
- LEVY, M.R. (1983). The methodology and performance of election day polls. *Public Opinion Quarterly*, 47, 54-67.
- MITOFSKY, W.J. (1991). A short history of exit polls. In *Polling in Presidential Election Coverage*. (Eds. P. Lavrakas and J. Holley). Newbury Park, California: Sage.
- MITOFSKY, W.J., and WAKSBERG, J. (1989). CBS models for election night estimation. Paper presented at American Statistical Association San Diego Winter Conference.
- MURRY, J.P., LASTOVICKA, J.L., and BHALLA, G. (1989). Demographic and lifestyle selection error in mall-intercept data. *Journal of Advertising Research*, 46-52.
- O'DAY, J., and WOLFE, A.C. (1984). Seat Belt Observations in Michigan - August/September 1983. Ann Arbor, Michigan: University of Michigan Transportation Research Institute.
- SUDMAN, S. (1980). Improving the quality of shopping center sampling. *Journal of Marketing Research*, 17, 423-31.
- STATISTICS CANADA (1979). Data Collection and Dissemination Methods for International Travel Statistics in Canada. International Travel Section, Statistics Canada.
- UNITED STATES TRAVEL AND TOURISM ADMINISTRATION (1989). In-flight Survey: Overseas and Mexican Visitors to the United States. Survey Period: January-December 1988. Washington, D.C.: United States Travel and Tourism Administration.
- WELLS, J.K., WILLIAMS, A.F., and LUND, A.K. (1990). Seat belt use on interstate highways. *American Journal of Public Health*, 80, 741-742.
- WOLFE, A.C. (1986). 1986 U.S. National Roadside Breathtesting Survey: Procedures and Results. Ann Arbor, Michigan: Mid-America Research Institute.
- WOOD, D. (1978). The Eastleigh Carrefour: a hypermarket and its effects. London: U.K. Department of the Environment.
- YATES, F. (1981). *Sampling Methods for Censuses and Surveys* (4th ed.). London: Charles Griffin.
- ZIEGLER, P.N. (1983). Guidelines for Conducting a Survey of the Use of Safety Belts and Child Safety Seats. Washington, D.C.: U.S. Department of Transportation.

## **A Sampling and Estimation Methodology for Sub-Annual Business Surveys**

**M.A. HIDIROGLOU, G.H. CHOUDHRY and P. LAVALLÉE<sup>1</sup>**

### **ABSTRACT**

A sample design for the initial selection, sample rotation and updating for sub-annual business surveys is proposed. The sample design is a stratified clustered design, with the stratification being carried out on the basis of industry, geography and size. Sample rotation of the sample units is carried out under time-in and time-out constraints. Updating is with respect to the selection of births (new businesses), removal of deaths (defunct businesses) and implementation of changes in the classification variables used for stratification, *i.e.* industry, geography and size. A number of alternate estimators, including the simple expansion estimator and Mickey's (1959) unbiased ratio-type estimator have been evaluated for this design in an empirical study under various survey conditions. The problem of variance estimation has also been considered using the Taylor linearization method and the jackknife technique.

**KEY WORDS:** Continuous surveys; Sample updating; Ratio estimator; Variance estimation.

### **1. INTRODUCTION**

The universe for sub-annual business surveys continually changes on account of births, deaths, splits, mergers, amalgamations, and classification changes. The sample design associated with such a universe should have the following characteristics. Firstly, it should result in samples which reflect the changing structure of the population. Secondly, it should distribute response burden by rotating units in and out of the sample. Thirdly, if there are significant changes in the stratification of the universe, it should be possible to redraw a new sample which reflects the stratification and possible changes in sampling fractions. The resulting new sample should have maximum overlap with the previous sample in order to minimize abrupt changes in the estimates and increased costs due to the introduction of new units in the sample. The sample design which has been proposed to satisfy these requirements is that of a simple random sample of randomly formed rotation groups (clusters) within each of the strata. Each rotation group represents either a group of units or a single unit. All units within a selected rotation group are selected in the sample. Rotation of the sample takes place under the constraints that units must stay in the sample for a certain period of time and be kept out of the sample for at least a certain period of time after they have rotated out of the sample.

For given domains of interest, unbiased (or nearly unbiased) estimates are developed along with the associated measures of reliability (coefficients of variation). A desirable property of the estimation is that the estimates of domain totals should add up to the population total when the domains are exhaustive and non-overlapping. This can be ensured by using one set of weights which is independent of the domains.

In section 2, the rotation group sampling design is developed and a number of alternative estimation procedures are described in section 3. In section 4, the results of an empirical study showing the performance of these estimators under various survey conditions are given. Finally, section 5 contains some concluding remarks.

<sup>1</sup> M.A. Hidiroglou, G.H. Choudhry and P. Lavallée, Business Survey Methods Division, Statistics Canada, 11th floor R.H. Coats Building, Ottawa, Ontario, Canada, K1A 0T6.

## 2. SAMPLING DESIGN

### 2.1 Stratification and Sample Allocation

The stratification of a business universe is usually based on one or more of the following characteristics: industry, geography, and size. The size measure can be univariate (*e.g.* sales or number of employees) or multivariate (*e.g.* revenue and assets). In our context, the primary strata are cross-classifications of industry and geographic regions for which estimates are required. Within these primary strata, secondary strata are formed using the size measure of the units. The secondary strata are comprised of a completely enumerated “take-all” stratum and a number of strata called ‘take-some’ strata where sampling occurs. It is necessary to have a take-all stratum on account of the highly skewed nature of the business universe. The take-all stratum boundary can be determined by a method introduced by Hidiroglou (1986). This method finds the optimum boundary between the take-all and the take-some strata within each primary stratum so as to minimize the overall sample size for a given coefficient of variation. The determination of this boundary also takes into account that certain units are to be sampled with certainty irrespective of their size. These pre-specified “take-all” units are units which are to be included in the sample on account of their complex structures. An example of a unit with a complex structure could be one which operates in more than one of the primary strata. The boundaries for the take-some strata are obtained either using the cum  $\sqrt{f}$  rule introduced by Dalenius and Hodges (1959) or the cum  $\sqrt{x}$  rule given by Hansen *et al.* (1953). Here  $x$  is a size variable available for stratification of the units in the population.

The sample sizes for the primary strata are computed so as to satisfy planned levels of precision for certain key estimates. The computation of these sample sizes also takes into account the required allocation scheme of the units to the take-some strata. It is assumed that the information available for computing these sample sizes is well correlated with the planned key variables. Given that the take-all sample units have been taken into account, the remaining sample is allocated to the take-some strata within the primary stratum, proportional to  $M^q$  or  $X^q$ , where  $M$  is the number of units in the take-some stratum and  $X$  is the take-some stratum total for the size variable being considered. The power  $q$  where  $0 \leq q \leq 1$  is chosen according to the required allocation. Letting  $q = 1$  results in Neyman allocation, whereas as  $q$  approaches zero, the resulting coefficients of variation become more equal amongst the different strata provided that  $S_h/\bar{X}_h$  does not vary significantly from stratum to stratum and that the finite population correction factors can be ignored. The advantages of these power allocations are discussed in Bankier (1988). The allocation can be adjusted to achieve the desired minimum sample sizes and/or maximum weights for each secondary stratum.

The reliability criteria (in terms of coefficients of variation) can be associated with the primary strata in one of two ways. Either they can be specified for each primary stratum, or, for a given global (national) coefficient of variation (c.v.), the c.v. at the primary stratum level can be determined so that the c.v.’s for each industry group and geographic region are equal. An iterative procedure is used to determine the desired c.v.’s for each of the primary strata and hence the sample size within each primary stratum, so that the planned c.v.’s at the global and marginal levels are achieved.

### 2.2 Sampling Scheme

For each stratum, the  $M$  population units within that stratum are randomly allocated to a predetermined number  $P$  of population rotation groups, so that initially, the number of units in each of any two rotation groups differ by at most one unit. The number of rotation groups



is a function of sampling fractions, and time-in and time-out constraints. It may be noted that in order to achieve unbiasedness, the time-in and time-out constraints may sometimes have to be violated. A simple random (SRS) sample of  $p$  rotation groups is selected from the  $P$  population rotation groups. The number of rotation groups  $p$  to be selected is determined such that  $p/P$  is approximately equal to the desired sampling fraction  $f$ . The sample consists of all the units in the  $p$  selected rotation groups. Rotation of the sample occurs by acquiring an out-of-sample rotation group and dropping an in-sample rotation group. Births are randomly allocated to the  $P$  population rotation groups, one at a time, in a systematic fashion. Deaths are removed from the stratum only if they are detected by a source independent of the survey, or if they have been dead for more than a pre-specified period of time. Methods proposed by Kish and Scott (1971) are adopted for sample updating with re-stratification due to population units changing strata. The sample update maximizes the overlap between the current and the new samples. There are obvious advantages to redrawing the sample in this fashion. First, it minimizes the introduction of new units into the sample, resulting in a smoother transition from an operational point of view, and also minimizes cost. Second, discontinuity in the estimates on account of sample redraw is kept to a minimum.

There are other sampling schemes which can be used to select the sample and rotate the units. These include Poisson and collocated sampling. The properties of these schemes have been discussed by Brewer, Early and Joyce (1972), and by Sunter (1977). Poisson sampling as defined by Hajék (1964) allows each unit in the population to be drawn in the sample independently with a given probability of inclusion. Decisions as to whether the unit is selected in the sample or not are made using an independent random draw or Bernoulli trial for each unit. Supposing that the inclusion probability of a given unit  $i$  is  $\pi_i$ , and that a random number  $u_i$  uniformly distributed over the interval  $(0,1)$  is generated, then the  $i$ -th unit is selected if  $u_i \leq \pi_i$ . This probability of inclusion corresponds to the sampling fraction of the stratum that the unit belongs to. Although the advantage of Poisson sampling lies in the simple manner in which sample rotation is exercised, it has certain disadvantages. Its main disadvantage is that the realized sample size is a random variable. This can be serious if the number of units in the stratum is small, possibly resulting in samples of size zero. Early and Brewer (1971) remedied this weakness by using a scheme known as collocated sampling. Collocated sampling is similar to Poisson sampling but reduces the variation in sample size by equispacing, at the cell level, the units over the interval  $(0,1)$ . Properties of this method are provided in more detail in Brewer, Early and Hanif (1984). Whereas in Poisson sampling, the addition of births and removal of deaths do not affect the random numbers attached to existing units, the use of collocated sampling requires that these random numbers be slightly perturbed, possibly disturbing the rotation scheme by violating the time-in and time-out constraints.

The rotation group sampling scheme has several advantages over the two previously mentioned schemes. For the rotation group sampling scheme, in contrast to the Poisson scheme, the expected number of units on each rotation cycle is almost equal. The removal of dead units on a universal basis may disturb the balance of units amongst the different rotation groups. This can be remedied by periodically redrawing the sample with maximum overlap, keeping the stratification and sampling fractions unchanged. The rotation for the rotation group scheme can be performed without perturbing the units, thereby satisfying the time-in and time-out constraints. This may not necessarily be true with collocated sampling on account of the slight perturbations of the random numbers due to population births and deaths. These effects may become non-trivial over a long period of time. Another advantage of the rotation group scheme over the other two methods is that re-stratification and new sampling fractions can easily be accommodated while maximizing sample overlap.

### 2.2.1 Determination of the Number of Rotation Groups

Assume that for a given take-some stratum, the number of population units is  $M$  and that the desired sampling fraction is  $f$ . Let  $t_{in}$  be the desired number of occasions a unit should stay in the sample. Let  $t_{out}$  be the minimum required number of occasions a unit must stay out of the sample, once it has rotated out of the sample. The required number of population rotation groups “ $P$ ” and in-sample rotation groups “ $p$ ” are determined as follows. Let,  $x = \text{int} [t_{in} (1 - f)/f + 0.5]$  where  $\text{int}[\cdot]$  denotes the integer portion of the argument. Two conditions arise:

- If  $x \geq t_{out}$ , then the number of in-sample rotation groups is  $p = t_{in}$  and the number of population rotation groups is  $P = t_{in} + x$ .
- If  $x < t_{out}$ , then the number of in-sample rotation groups is

$$p = \text{int} \left[ \frac{f}{1 - f} t_{out} + 0.5 \right]$$

and the number of population rotation groups is  $P = p + t_{out}$ .

It must be noted that  $p/P$  is only approximately equal to  $f$  on account of the integer operations.

### 2.2.2 Allocation of Units to Rotation Groups

Given that at the time of initial selection, there are  $M$  population units to be allocated to  $P$  population rotation groups, two distinct cases arise with respect to the relative sizes of  $M$  and  $P$ :  $M \geq P$  or  $M < P$ .

When  $M \geq P$ , at least one unit can be allocated to each population rotation group. Suppose  $M = aP + \ell$ , where  $a > 0$  and  $\ell \geq 0$  are integers. In order to equalize the rotation group sizes as much as possible at the time of initial selection and on subsequent occasions, the following procedure is used. A 2 by  $P$  matrix is used to assign a rotation sequence to the units that will satisfy the requirements of almost equal rotation group sizes. It is used for initial sample selection and subsequent addition of births. The first “assignment” row is labelled from 1 to  $P$ , whereas the second “rotation” row is a randomized order of the first row. The corresponding rotation group numbers in the second row determine which units are in sample at any point in time. The  $M$  population units are assigned sequentially to the assignment rotation group numbers 1, 2, ...,  $P$ , the  $P$ -th unit going to the  $P$ -th assignment rotation group number. The  $(P + 1)$ -th unit is assigned to assignment rotation group number 1 and so on. This eventually results in having the first “ $\ell$ ” assignment rotation groups with  $(a + 1)$  units and the next  $(P - \ell)$  assignment rotation groups with  $(a)$  units. The rotation group to which the  $M$ -th unit is assigned is termed the last assignment rotation group. This rotation group, which is assigned rotation group number at time of initial selection, is used for assigning future births starting from the next assignment rotation group number, *i.e.*  $\ell + 1$ .

When  $M < P$ , the  $M$  population units can only be allocated to a subset of  $M$  out of the  $P$  rotation groups. These rotation groups must be as equispaced as possible to ensure that the expected sample size,  $\bar{m} = fM$ , will be achieved from one survey occasion to the next. For this case the allocation matrix is 2 by  $M$ . The first assignment row is labelled from 1 to  $M$ . The second rotation row is a randomization of  $M$  “ $z$ ” numbers where  $1 \leq z_i < z_j \leq P$  for  $i \neq j, i = 1, \dots, M$  and  $j = 1, \dots, M$ . The “ $z$ ” numbers are created as follows.

- Find integers  $s$  and  $q$  such that  $P = sM + q$  where  $q < M$  and  $s \geq 0$ .
- Generate  $r_j$  ( $j = 1, \dots, M$ ) numbers randomly assuming the values 0 or 1, such that  $q$  of them have the value equal to 1 and  $M - q$  of them have the value equal to 0.

- iii) Select a random integer “ $b$ ” such that  $1 \leq b \leq P$ .
- iv) Compute  $z_1 = (b + r_1 - 1) \bmod P + 1$  and  $z_j = (z_{j-1} + s + r_j - 1) \bmod P + 1$  for  $j = 2, \dots, M$ .
- v) Randomize the “ $z$ ” numbers. Let the sequence of randomized “ $z$ ” numbers be  $z_{i_1}, z_{i_2}, \dots, z_{i_M}$ .

Now the  $M$  population units are assigned sequentially to the  $M$  assignment rotation group numbers, thereby picking up their rotation numbers. The last assignment rotation group number is  $M$ . Future births will be assigned starting from assignment rotation group number 1.

It is now a simple matter to perform the basic functions of sample selection and updating.

### 2.2.3 Sample Selection and Updating

At time of initial sample selection, a given stratum will have  $N = \min(M, P)$  distinct rotation groups. The units belonging to the initial sample are those whose rotation numbers are included in the closed sampling interval  $[1, p]$ . When  $M \geq P$ , the number of in-sample rotation groups  $n$  is equal to  $p$ . When  $M < P$ , the number of in-sample rotation groups  $n$  is approximately equal to  $fN$  on account of the equispacing.

Sample rotation is carried out by shifting the sampling interval by one rotation group at each sampling occasion in a circular fashion. On the  $t$ -th occasion, units in the sample are those whose rotation number is contained in the interval defined as

- i)  $[(t - 1) \bmod P + 1, (t + p - 2) \bmod P + 1]$ , if  $(t - 1) \bmod P \leq (P - p)$  and
- ii)  $[1, (p - P) + (t - 1) \bmod P] \cup [(t - 1) \bmod P + 1, P]$ , otherwise.

Effectively, rotation occurs by dropping a rotation group from in-sample and acquiring a rotation group from out-of-sample in a modular fashion.

“Births” occur as a result of starting a new business activity, or a change of industrial activity of a unit from out-of-scope to in-scope for the survey. Births are stratified and given an assignment rotation group number within the stratum as follows. Assuming the last assignment rotation group number was  $\ell$ , where  $1 \leq \ell \leq P$ , the  $q$ -th birth will be given the assignment rotation group number  $(\ell + q) \bmod P$ . The next birth will be given the assignment rotation group number  $(\ell + q + 1) \bmod P$ . The rotation number is then immediately obtained through the one-to-one correspondence between the assignment and rotation numbers.

“Deaths” occur as a result of the termination of business activity for in-scope units or changes of industrial activity from in-scope to out-of-scope to the survey. Deaths that occur in a take-all stratum are immediately removed from the population and sample. Deaths that are part of a take-some stratum are removed immediately if they are identified as such by a source independent of the survey process. Otherwise, they are removed after a given time period. This time period should be sufficiently long so that most of the population deaths would have been identified. Deaths in the sample and in this latter category which have not yet been removed are assigned a value of zero for estimation purposes. Classification values are also retained as such until they have been identified as changes by a source independent of the survey.

### 2.2.4 Periodic Resampling

The sampling frame changes continually due not only to births and deaths, but also due to changes of classification variables used in the stratification (*i.e.*, geography, industry and size). These changes in the classification variables are reflected in the estimation process by



use of domain estimation (*i.e.* estimation for sub-populations). That is, the latest classification is assigned to data for tabulation purposes, using the original sampling weight. Over a period of time, changes in classification may be sufficiently important to require the examination of the stratification and subsequent sampling rates. One solution would be to redraw an independent sample, taking into account these changes, but ignoring the current sample. Such an approach has certain disadvantages from an operational point of view. An independent redraw implies that i) the newly sampled units must be initiated into the sample, ii) time-in and time-out constraints can be violated, and iii) the estimates may change substantially. It is therefore desirable to maximize the overlap between the current sample and the new sample. The following methodology provides such a procedure for resampling. It is an adaptation of the Kish and Scott (1971) method, and is based on the property that each rotation group is a simple random sample from the population rotation groups.

At time of resampling, rotation will have occurred at different rates amongst the strata, resulting in sampling intervals with different starting and end points. Hence, assuming that rotation started at time  $t_1$  and that we are currently at time  $t_2$ , the number of rotations that have occurred is  $r = t_2 - t_1 + 1$ . At time  $t_2$ , the sampling interval(s) associated with a given stratum currently labelled as  $k$  ( $k = 1, 2, \dots, K$ ) is(are)

$$[(r - 1) \bmod P_k + 1, (r + p_k - 2) \bmod P_k + 1] \quad \text{if} \quad (r - 1) \bmod P_k \leq (P_k - p_k)$$

and

$$[1, p_k - P_k + (r - 1) \bmod P_k] \quad \text{and} \quad [(r - 1) \bmod P_k + 1, P_k] \quad \text{otherwise.}$$

The first step associated with the resampling is to relabel the different sampling intervals, which have different starting points, into sampling intervals which have the same starting point. For the  $k$ -th stratum, the resulting sampling interval is  $[1, p_k]$ . Let  $b$  denote the starting point of the sampling interval at time  $t_2$  where  $b$  is given by  $(r - 1) \bmod P_k + 1$ . All units labelled with rotation number “ $g$ ” are relabelled as  $(g - b + 1)$  if  $b \leq g \leq P_k$  and as  $P_k - (b - g - 1)$  otherwise.

The second step is to associate with each population unit currently classified to stratum  $k$  its new stratum “ $h$ ”. The population units of the new  $h$ -th stratum,  $U_h$ , can therefore be expressed as the union of  $K$  non-overlapping and exhaustive sets  $U_{hk}$ ,  $h = 1, 2, \dots, L$ . Each set  $U_{hk}$  is comprised of population units whose new stratification is  $h$  and current stratification is  $k$ . Some of these sets may be empty.

The third step is to rank, on the 0 to 1 scale, sampling units within each set  $U_{hk}$ , taking into account their current rotation numbers. Assume that there are  $M_{hk}$  units in the set  $U_{hk}$  and that their current rotation numbers are labelled between 1 and  $P_k$ . Rank these units from 1 to  $M_{hk}$  based on their associated current rotation number. Units which have the lowest rotation numbers are assigned the lowest ranks and units which have the highest rotation numbers are assigned the highest ranks. If there are any ties, these can be broken up randomly by generating uniform random numbers. This results in the units in set  $U_{hk}$  to be ranked from 1 to  $M_{hk}$ . Next, a unit with rank “ $i$ ” in set  $U_{hk}$ ,  $1 \leq i \leq M_{hk}$ , is assigned a number  $r_{hki} = (a_{hk} + i - 1)/M_{hk}$ , where  $a_{hk}$  is a uniformly generated random number between 0 and 1 for each set  $U_{hk}$  within  $U_h$ . These numbers represent the current rotation groups transformed to the range 0 and 1. Assume that the new sampling fraction associated with the new stratum is  $f_h$  and that the current sampling fraction is  $f_k$ . If  $f_h \geq f_k$ , this implies that all units currently sampled in  $U_{hk}$  will stay in the new sample and that units in the closed interval  $[0, f_h]$  will be included in the new sample. If  $f_h < f_k$ , this implies that units must be dropped (rotated out) from the current

sample. The units which must be dropped are those which have the lowest  $r_{hki}$  values. These represent the rotation groups which have been in the sample the longest. In order that the units in the new sample be contained in the closed interval  $[0, f_h]$  it is necessary to relabel the  $r_{hki}$ 's as  $r_{hki} - (f_k - f_h)$  if  $r_{hki} \geq (f_k - f_h)$  and as  $r_{hki} - (f_k - f_h) + 1$  otherwise. Assuming that the population units belonging to the new  $h$ -th stratum are ranked based on the ordered  $r_{hki}$ 's, define  $b_{hi} = i / (M_h + 1)$ ,  $i = 1, 2, \dots, M_h$ . Using the  $b_{hi}$ 's, new rotation numbers will be obtained as follows. For a given new stratum  $h$ , let  $N_h$  be the number of distinct rotation groups. Form  $N_h$  disjoint intervals

$$I_u = \begin{cases} [(u-1)/N_h, u/N_h] & \text{for } u=1, \dots, N_h-1 \\ [(N_h-1)/N_h, 1] & \text{for } u = N_h. \end{cases}$$

The union of these intervals is the closed interval  $[0,1]$   $D_{u_i}$ . For  $D_{u_j}$  the new stratum  $h$ , label the new rotation numbers as where  $D_1, D_2, \dots, D_{N_h}$  where  $D_{u_i} < D_{u_j}$  for  $u_i < u_j$ ,  $u_i = 1, \dots, N_h$ . The  $i$ -th unit acquires rotation number  $D_u$  if its corresponding  $b_{hi}$  value belongs to the interval  $I_u$ . Assuming that all the  $M_h$  units have been assigned new rotation numbers in this fashion, the units in sample will be those whose rotation number belongs to the interval  $[1, p_h]$ .

### 3. WEIGHTING AND ESTIMATION

The simplest estimator which can be used in conjunction with the rotation group design described in Section 2.2 is the simple expansion (or simple domain) estimator. Although this estimator is unconditionally unbiased, it can have a large conditional bias when the rotation group sizes are not balanced. The removal of dead units can cause such an imbalance in the distribution of rotation group sizes. Other estimators which take the auxiliary rotation group size information into account have therefore been considered. These include the separate and combined ratio estimators. A drawback of the separate estimator is that its bias may accumulate in a non-trivial manner across strata. The combined estimator will have negligible bias, but possibly large variances for stratum level estimates. We have therefore evaluated the performance of an unbiased separate ratio estimator due to Mickey (1959). The penalty for achieving unbiasedness is an increase in the variance. The primary objective is to determine which of the above estimators is the most suitable one for the rotation group design. The criteria for choosing the most appropriate estimator will be based on bias and mean squared error. In order to simplify the comparisons, it will be assumed that each sampled unit has valid response data.

As mentioned earlier, the  $h$ -th stratum ( $h = 1, 2, \dots, L$ ) is defined at some given level of industry, geography and size. Estimates are required for domains which can span all the sampling strata or be a subset of these strata. Examples of such domains are aggregations of variables of interest at the sub-provincial level given that the sampling may have occurred at a higher level, e.g. province. A desirable feature of the estimates is that the sum of any non-overlapping domain set must always add up to the domain defined as their union. In order to achieve consistency, only one set of weights can be used.

Let  $y$  denote the characteristic of interest and  $y_{hij}$  be its value for the  $j$ -th unit in rotation group (cluster)  $i$  of stratum  $h$ . Let  $\delta_{hij}(d)$  be an indicator variable defined as 1 if the  $hij$ -th unit belongs to domain "d", and 0 otherwise. Then, the parameter of interest is the population total  $Y(d)$  given by:

$$Y(d) = \sum_{h=1}^L \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} y_{hij}(d),$$

where  $y_{hij}(d) = \delta_{hij}(d) y_{hij}$ .

As described earlier, we have a simple random sample of  $n_h$  rotation groups selected without replacement from the  $N_h$  rotation groups in the  $h$ -th stratum. Let  $M_{hi}$  be the number of units in the  $i$ -th sampled rotation group within stratum  $h$ . Without loss of generality, we can assume that the sampled rotation groups are indexed  $i = 1, 2, \dots, n_h$ . Let  $y_{hi}(d)$  be the total response of the units belonging to domain “ $d$ ” from the  $i$ -th sampled rotation group within stratum  $h$ , *i.e.*

$$y_{hi}(d) = \sum_{j=1}^{M_{hi}} y_{hij}(d), i = 1, 2, \dots, n_h.$$

We will consider a number of alternative estimators for the population parameter  $Y(d)$  and their corresponding variance. The estimators considered are of the form,

$$\hat{Y}_h(d) = \sum_{i=1}^{n_h} w_{hi} y_{hi}(d),$$

where  $w_{hi}$  is the product of the design weight and an adjustment which reflects the estimation procedure used. Estimators of  $Y(d)$  are obtained by aggregating over strata, that is,

$$\hat{Y}(d) = \sum_{h=1}^L \hat{Y}_h(d).$$

### 3.1 Estimators of Total

#### A. Simple Expansion Estimator

Since the probability of selecting a rotation group in the  $h$ -th stratum is  $n_h/N_h$ , the design weight is  $w_{hi} = N_h/n_h$  for  $i = 1, 2, \dots, n_h$ ,  $h = 1, 2, \dots, L$ . The simple expansion estimator is given by

$$\hat{Y}_E(d) = \sum_{h=1}^L N_h \bar{y}_h(d), \quad (3.1)$$

where

$$\bar{y}_h(d) = n_h^{-1} \sum_{i=1}^{n_h} y_{hi}(d).$$

As mentioned earlier, this estimator is unconditionally unbiased, but it can have a large conditional bias. Moreover, it may not be very efficient because it does not make use of available auxiliary information, such as rotation group sizes. As the variation in the rotation group sizes may increase over time on account of removal of deaths, it may become more and more inefficient.

#### B. Separate Ratio Estimator

If the correlation between  $y_{hi}(d)$  and rotation group sizes  $M_{hi}$  is large, efficiency gains can be realized through the separate ratio estimator defined as

$$\hat{Y}_{SR}(d) = \sum_{h=1}^L \left( \frac{M_h}{\bar{m}_h} \right) \bar{y}_h(d) \quad (3.2)$$



where

$$\bar{m}_h = n_h^{-1} \sum_{i=1}^{n_h} M_{hi}$$

and

$$M_h = \sum_{i=1}^{N_h} M_{hi}.$$

One major drawback of this estimator is that it is subject to the ratio estimation bias. Consequently, if the bias tends to be positive or negative in the majority of the strata, its accumulated effect can be quite significant when aggregating over the strata.

C. Combined Ratio Estimator

The accumulated effect of aggregation bias can be significantly reduced using a combined version of the ratio estimator. The combined ratio estimator is given by

$$\hat{Y}_{CR}(d) = M \frac{\sum_{h=1}^L N_h y_h(d)}{\sum_{h=1}^L N_h \bar{m}_h}, \tag{3.3}$$

where  $M = \sum_{h=1}^L M_h$ .

D. Unbiased Ratio-type Estimator

The bias problem caused by the ratio estimation can be completely eliminated using the following adjusted ratio-type estimator suggested by Mickey (1959). The Mickey estimator is given by

$$\hat{Y}_{MI}(d) = \sum_{h=1}^L \left( \bar{r}_h(d) M_h + (N_h - n_h + 1) \left[ \sum_{i=1}^{n_h} y_{hi}(d) - m_h \bar{r}_h(d) \right] \right), \tag{3.4}$$

where

$$\bar{r}_h(d) = \frac{1}{n_h} \sum_{j=1}^{n_h} \bar{r}_h^{(j)}(d); \bar{r}_h^{(j)}(d) = \frac{\sum_{i \neq (j)} y_{hi}(d)}{\sum_{i \neq (j)} M_{hi}}; m_h = \sum_{i=1}^{n_h} M_{hi}.$$

An undesirable feature of the Mickey estimator is that it can have weights less than one, including negative weights.

For the separate and combined ratio estimators, the variances are estimated using the Taylor linearization method. In the case of Mickey's estimator, a jackknife procedure is used, leaving out one rotation group at a time and re-computing Mickey's estimator for the remaining  $(n_h - 1)$  rotation groups in the sample. Denote each jackknifed estimator as for  $\hat{Y}_{MI,h}^{(j)}(d)$  for  $j = 1, 2, \dots, n_h$ ,

where

$$\hat{Y}_{MI,h}^{(j)}(d) = \sum_{i \neq (j)} w_{hi}^{(j)} y_{hi}(d)$$

with

$$w_{hi}^{(j)} = [M_h - (m_h - M_{hj})] (N_h - n_h + 2) b_{hi}^{(j)} + (N_h - n_h + 2)$$

and

$$b_{hi}^{(j)} = (n_h - 1)^{-1} \sum_{i \neq (j)} \frac{1}{(m_h - M_{hj} - M_{hi})}.$$

A jackknife variance estimator of  $\hat{Y}_{MI,h}(d)$  is given by

$$v_j(\hat{Y}_{MI,h}(d)) = (1 - f_h) \frac{(n_h - 1)}{n_h} \sum_{j=1}^{n_h} (z_h^{(j)}(d) - \bar{z}_h(d))^2,$$

where  $z_h^{(j)}(d) = \hat{Y}_{MI,h}^{(j)}(d)$  and  $\bar{z}_h(d) = n_h^{-1} \sum_{j=1}^{n_h} z_h^{(j)}(d)$ .

It can be shown that all the estimators are equivalent and unconditionally unbiased when the rotation group sizes  $M_{hi}$  are all equal in each stratum  $h$ . However once the rotation group sizes ( $M_{hi}$ ) become unequal, all estimators, except for the simple expansion and the Mickey estimator, are unconditionally biased. For these estimators, the magnitude of their unconditional biases and their efficiency was assessed in a simulation study which is presented next.

#### 4. SIMULATION STUDY

The purpose of this simulation was to determine which of the four estimators of aggregate total  $Y(d)$  and the stratum total  $Y_h(d)$  would be the most “appropriate” for the sample design described in Section 2. For simplicity, the simulations were confined to a single variable ( $y$ ), gross business income (GBI). Also, for the purpose of this simulation the domains coincided with strata. Therefore, the symbol “ $d$ ” used to denote the domain will be omitted.

##### 4.1 Description of the Study

The universe for the simulation study was defined as the set of smaller sized units belonging to the Wholesale Trade sector in the province of Québec for the May 1989 reference period. The size of each unit was based on a GBI derived from payroll deductions using a ratio model. Units whose GBI was below a given threshold were retained, resulting in a population of 10,953 units. The stratification of this population was defined on the basis of Standard Industrial Classification at the 3 digit level. This resulted in 30 strata with a minimum stratum size of 18 units. For each of the 30 strata, 16 rotation groups were formed by randomly assigning the units to the rotation groups as described in Section 2.2.2.

For each stratum  $h$ , samples of 4 rotation groups were obtained from the 16 rotation groups using simple random sampling without replacement. From each stratum there were 1,820 possible samples of size 4. Over the 30 strata there were 54,600 (30 strata times 1,820 samples per stratum) possible different estimates for the separate ratio estimation procedure. On the other hand, for the combined ratio estimator, a total of  $(1,820)^{30}$  different estimates could be produced. For the simple expansion, the separate ratio estimator, and Mickey’s estimator, all 54,600 possible samples were drawn. For the combined ratio estimator, 100,000 samples were randomly drawn from the  $(1,820)^{30}$  possible samples.

##### 4.2 Evaluation Criteria

The evaluation criteria involved bias and mean squared error. These are described next.

For each selected sample  $k$ , an estimate  $\hat{Y}_h^{(k)}$  was produced for each stratum  $h$  and for each of the four estimators. The stratum expectation  $E(\hat{Y}_h^{(k)})$  of this estimate was obtained as

$$E(\hat{Y}_h) = \frac{1}{K} \sum_{k=1}^K \hat{Y}_h^{(k)},$$

where  $K$  is the total number of samples drawn. It should be noted that for estimators (3.1) – (3.2) and (3.4),  $E(\hat{Y}_h)$  was in fact the true expectation since all possible samples were drawn. For the combined ratio estimator (3.3), it corresponded to an unbiased estimate of the expectation. The resulting stratum bias was

$$\text{Bias}(\hat{Y}_h) \doteq E(\hat{Y}_h) - Y_h.$$

The total bias,  $\text{Bias}(\hat{Y})$ , was obtained by summing the stratum bias over all strata. For estimators (3.1) – (3.2) and (3.4), we have that

$$\text{Var}(\hat{Y}_h) \doteq \frac{1}{K} \sum_{k=1}^K (\hat{Y}_h^{(k)} - E(\hat{Y}_h))^2$$

and

$$\text{Var}(\hat{Y}) = \sum_{h=1}^L \text{Var}(\hat{Y}_h).$$

For the combined ratio estimator (3.3), we have that

$$\text{Var}(\hat{Y}_h) \doteq \frac{1}{K-1} \sum_{k=1}^K (\hat{Y}_h^{(k)} - E(\hat{Y}_h))^2$$

and

$$\text{Var}(\hat{Y}) = \frac{1}{K-1} \sum_{k=1}^K (\hat{Y}^{(k)} - E(\hat{Y}))^2,$$

where  $\hat{Y}^{(k)} = \sum_{h=1}^L \hat{Y}_h^{(k)}$  and  $E(\hat{Y}) = \sum_{h=1}^L E(\hat{Y}_h)$ .

Finally, the stratum mean squared error,  $\text{MSE}(\hat{Y}_h)$ , of each estimator was defined as

$$\text{MSE}(\hat{Y}_h) = \text{Var}(\hat{Y}_h) + (\text{Bias}(\hat{Y}_h))^2$$

while the aggregate mean squared error,  $\text{MSE}(\hat{Y})$ , of each estimator was given by

$$\text{MSE}(\hat{Y}) = \text{Var}(\hat{Y}) + (\text{Bias}(\hat{Y}))^2.$$

Four criteria were used in comparing the relative behaviour of the proposed estimators. The first criterion was absolute relative bias. The stratum average absolute relative bias was computed as

$$\overline{\text{ARB}} = \frac{1}{L} \sum_{h=1}^L \left| \text{Bias}(\hat{Y}_h) \right| / Y_h,$$

while the aggregate absolute relative bias was computed as

$$\text{ARB} = \left| \sum_{h=1}^L \text{Bias}(\hat{Y}_h) \right| / Y,$$



where

$$Y = \sum_{h=1}^L Y_h.$$

The second criterion was the ratio of absolute bias to standard error which was called “absolute standard bias”. The stratum average absolute standard bias was computed as

$$\overline{\text{ASB}} = \frac{1}{L} \sum_{h=1}^L \left| \text{Bias}(\hat{Y}_h) \right| / \sqrt{\text{Var}(\hat{Y}_h)}$$

while at the aggregate level, it was computed as

$$\text{ASB} = | \text{Bias}(\hat{Y}) | / \sqrt{\text{Var}(\hat{Y})}.$$

Following Cochran (1977), a reasonable value for the maximum acceptable bias over the standard error should not exceed 10%. Indeed, since the precision of an estimator is usually measured by its variance and not by its MSE, too large a bias as compared to the standard deviation would give a false impression of the precision of the estimator used.

The third criterion was efficiency, defined as the ratio of the root mean squared error of the estimator under study,  $\text{RMSE}(\hat{Y}^{EST})$ , to that of the simple expansion estimator  $\text{RMSE}(\hat{Y}^{EXP})$ . The stratum average relative efficiency was computed as

$$\overline{\text{EFF}} = \frac{1}{L} \sum_{h=1}^L \{ \text{RMSE}(\hat{Y}_h^{EXP}) / \text{RMSE}(\hat{Y}_h^{EST}) \},$$

while at the aggregate level, the relative efficiency was computed as

$$\text{EFF} = \text{RMSE}(\hat{Y}^{EXP}) / \text{RMSE}(\hat{Y}^{EST}).$$

Finally, the fourth criterion was to observe the proportion of negative weights.

### 4.3 Description of the Scenarios

Four different scenarios were considered for the possible configuration of the population of rotation groups for the rotation group sample design described in Section 2.2. The four scenarios provided different combinations of the rotation group size balance (good, poor) and of the correlation between the rotation group sizes  $M_{hi}$  and the survey variable  $y_{hi}$  (good, scattered). In the context of rotation group balance, “good” means that the rotation groups do not differ much in size, whereas “poor” means that they differ significantly. In the context of correlation, “good” means that the correlation between the survey variable and the rotation group size is quite high throughout the strata, whereas “scattered” means that it varies from low to high amongst the strata.

These scenarios represent possible configurations that will arise as the survey progresses through time. Scenario 1 reflects the survey at time of initial selection: for this case, the balance of rotation group sizes is good, and the correlation between rotation group sizes and the survey variable is good. Scenario 2 reflects the deterioration of the correlation (scattered) between the rotation group size and the survey variable as time progresses, due to dead units accumulating in the population. For this scenario, since the dead units have not been removed from the population, the balance in rotation group sizes is good, but the correlation between the survey

variable and the rotation group size is weakened. Scenario 3 implies that removal of the dead population units may result in imbalance of the rotation group sizes (poor), but strengthening the correlation (good) between rotation group size and the survey variable. Finally scenario 4 represents the worst possible case, which is poor correlation between rotation group size and the survey variable, and poor balance in rotation group sizes.

Scenario 1 was constructed by varying the rotation group sizes and leaving the GBI values  $y_{hi}$  unchanged for all the rotation groups. The 16 rotation group sizes were varied by sorting them in ascending order of  $y_{hi}$ . Their size was set as follows. For rotation groups 1-4, 5-8, 9-12 and 13-16, the rotation group size was set to  $0.22 M_h/4$ ,  $0.24 M_h/4$ ,  $0.26 M_h/4$  and  $0.28 M_h/4$  respectively. The average correlation between the GBI and the rotation group sizes was 0.86, ranging from 0.69 to 0.96 at the individual stratum level. The average coefficient of variation of the rotation group sizes was 9.2%.

For scenario 2, the population units were randomly permuted and assigned systematically to one of 16 rotation groups, using the procedure described in Section 2.2.2. Approximately 20% of the population units were then randomly assigned a  $y$ -value of zero to represent a high proportion of dead units. The overall correlation between the GBI and the rotation group sizes was 0.11, ranging from  $-0.23$  to  $0.74$  at the individual stratum level. The average coefficient of variation of the rotation group sizes was 4.1%.

For scenario 3 the procedure was similar to scenario 1 except that the rotation group sizes differed. For rotation groups 1-4, 5-8, 9-12 and 13-16, the rotation group size was set as  $0.05 M_h/4$ ,  $0.20 M_h/4$ ,  $0.30 M_h/4$  and  $0.45 M_h/4$  respectively. The overall correlation between the GBI and the rotation group sizes was 0.87, ranging from 0.70 to 0.96 at the stratum level. The average coefficient of variation of the rotation group sizes was 60.2%.

For scenario 4 a random rotation group size was assigned independently of the GBI values as follows. Suppose that for each stratum  $h$ ,  $a_h = \min\{M_{hi}; i = 1, \dots, N_h\}$  and  $b_h = \max\{M_{hi}; i = 1, \dots, N_h\}$ . For each stratum  $h$ , the size  $M_{hi}^*$  for rotation group  $i$  was set to  $r_h e_{hi}$  where  $e_{hi}$  is uniformly distributed on the interval  $(a_h, b_h)$ . Here  $r_h$  is a scaling factor such that  $M_h = \sum_{i=1}^{N_h} M_{hi}^*$ . The average correlation was 0, ranging from  $-0.49$  to  $0.56$  at the stratum level. The average coefficient of variation of the rotation group sizes was 49.2%.

#### 4.4 Discussion of Results

Based on the 4 scenarios described in the previous section, simulations were performed to compute the absolute relative bias (ARB), the absolute standard bias (ASB), the efficiency (EFF), and the proportion of weights less than or equal to 0. Those quantities were computed for each individual stratum and at the aggregate level. The results are given in Tables 1 to 3. Note that all of these results are presented as percentages.

In terms of absolute relative bias (ARB), as shown in Table 1, both the simple expansion and Mickey's estimator have no bias, as expected, neither at the overall nor at the stratum level. The separate ratio estimator displays the most absolute relative bias while the combined ratio estimator displays the least relative bias. For the biased estimators, the absolute relative bias increases as the coefficient of variation of the rotation group sizes increases, and the correlation between the rotation group sizes and the variable of interest decreases.

Turning to absolute standard bias (ASB), as shown in Table 2, the following observations can be made. The separate ratio estimator is unacceptable for most scenarios using this criterion. Its performance worsens as the variation in rotation group sizes increases, and as the correlation between the rotation group sizes and as the variable of interest decreases. The performance of the combined ratio estimator is acceptable, both at the aggregate and stratum level.

**Table 1**  
Percentage Absolute Relative Bias ( $\overline{ARB}$ )

Scenario	Aggregate Level		Stratum Level	
	Separate Ratio	Combined Ratio	Separate Ratio	Combined Ratio
1	1.27	0.07	1.31	0.11
2	0.02	0.01	0.24	0.05
3	2.88	0.14	3.19	0.29
4	5.51	0.22	5.72	0.30

**Table 2**  
Percentage Absolute Standard Bias ( $\overline{ASB}$ )

Scenario	Aggregate Level		Stratum Level	
	Separate Ratio	Combined Ratio	Separate Ratio	Separate Ratio
1	13.41	0.76	3.37	0.24
2	0.44	0.26	0.58	0.25
3	45.64	2.13	12.11	0.69
4	43.29	1.96	9.88	0.71

The behaviour of the estimators with respect to relative efficiency (EFF) is provided in Tables 3a and 3b. For Scenario 1, which represents good rotation group balance and good correlation, all the estimators are nearly equivalent, both at the aggregate and the stratum levels. For Scenario 2, which represents well balanced rotation groups and scattered correlation, the same conclusion holds. For Scenario 3, which represents poor rotation group balance and good correlation between the rotation group sizes and the survey variable, the ranking of the estimators at the aggregate level from highest EFF to lowest EFF is: i) the combined ratio, ii) the separate ratio estimator, iii) Mickey’s estimator, and iv) the simple expansion estimator. For Scenario 4, which represents the worst in terms of rotation group balance and correlation between the rotation group sizes and the survey variable, the best estimator at both the aggregate and stratum levels is the simple expansion estimator. The combined ratio estimate is the next best choice.

Weights smaller than zero occurred for the Mickey estimator in 2% of the cases.

In conclusion, given the above four scenarios, the combined ratio estimator is a reasonable choice for estimation for sub-annual surveys which use the rotation group design. The simple expansion estimator may also be considered on account of its simplicity. However, one should be aware of its poor conditional properties if the rotation group sizes are not balanced.



**Table 3a**  
Percentage Relative Efficiency (EFF) at the Aggregate Level

Scenario	Simple Expansion	Separate Ratio	Combined Ratio	Mickey
1	100.0	108.0	107.9	107.3
2	100.0	100.2	99.8	100.1
3	100.0	148.3	160.3	143.5
4	100.0	74.3	92.3	84.3

**Table 3b**  
Percentage Average Relative Efficiency ( $\overline{\text{EFF}}$ ) at the Stratum Level

Scenario	Simple Expansion	Separate Ratio	Combined Ratio	Mickey
1	100.0	109.6	108.6	108.6
2	100.0	100.9	99.5	100.6
3	100.0	183.3	180.2	174.2
4	100.0	80.0	99.4	83.7

5. CONCLUSION

In this paper, we have presented a sample design which can accommodate the necessary requirements for a sub-annual business survey. These requirements have included initial sample selection, sample rotation and updating. Given this rotation group design, a number of estimation procedures have been considered and they have been evaluated via a simulation study. These estimation procedures are equivalent when the rotation group sizes are well balanced within each of the strata. In the case of unbalanced rotation group sizes, the use of the combined ratio estimator which used rotation group sizes as auxiliary information is recommended.

ACKNOWLEDGEMENTS

The authors would like to thank Lyne Guertin for the programming of the simulation study. We would also like to thank the referees and Professor J.N.K. Rao for valuable comments and constructive suggestions.

## REFERENCES

- BANKIER, M.D. (1988). Power allocations: Determining sample sizes for subnational areas. *The American Statistician*, 42, 174-177.
- BREWER, K.R.W., EARLY, L.J., and HANIF, M. (1984). Poisson, modified Poisson and collocated sampling. *Journal of Statistical Planning and Inference*, 10, 15-30.
- BREWER, K.R.W., EARLY, L.J., and JOYCE S.F. (1974). Selecting several samples from a single population. *Australian Journal of Statistics*, 14, 232-239.
- COCHRAN, W.G. (1977). *Sampling Techniques*, (3<sup>rd</sup> Edition). New York: John Wiley.
- DALENIUS, T., and HODGES, J.L., Jr. (1959). Minimum variance stratification. *Journal of the American Statistical Association*, 54, 88-101.
- EARLY, L.J., and BREWER, K.R.W. (1971). Some estimators for arbitrary probability sampling. Master's thesis.
- HAJÉK, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics*, 35, 1491-1523.
- HANSEN, M.H., HURWITZ, W.N., and MADOW, W.G. (1953). *Sample Survey Methods and Theory*. New York: John Wiley and Sons.
- HIDIROGLOU, M.A. (1986). The construction of a self-representing stratum of large units in survey design. *The American Statistician*, 40, 27-31.
- KISH, L., and SCOTT, A. (1971). Retaining units after changing strata and probabilities. *Journal of the American Statistical Association*, 66, 461-470.
- MICKEY, M.R. (1959). Some finite population unbiased ratio and regression estimators. *Journal of the American Statistical Association*, 54, 594-612.
- RAO, J.N.K., and KUZIK, R.A. (1974). Sampling errors in ratio estimation. *Sankhyā. Series, C*, 36, 43-58.
- SUNTER, A.B. (1977). Response burden, sample rotation and classification renewal in economic surveys. *International Statistical Review*, 45, 209-222.

## County Estimates of Wheat Production

ELIZABETH A. STASNY, PREM K. GOEL  
and DEBORAH J. RUMSEY<sup>1</sup>

### ABSTRACT

Although farm surveys carried out by the USDA are used to estimate crop production at the state and national levels, small area estimates at the county level are more useful for local economic decision making. County estimates are also in demand by companies selling fertilizers, pesticides, crop insurance, and farm equipment. Individual states often conduct their own surveys to provide data for county estimates of farm production. Typically, these state surveys are not carried out using probability sampling methods. An additional complication is that states impose the constraint that the sum of county estimates of crop production for all counties in a state be equal to the USDA estimate for that state. Thus, standard small area estimation procedures are not directly applicable to this problem. In this paper, we consider using regression models for obtaining county estimates of wheat production in Kansas. We describe a simulation study comparing the resulting estimates to those obtained using two standard small area estimators: the synthetic and direct estimators. We also compare several strategies for scaling the initial estimates so that they agree with the USDA estimate of the state production total.

**KEY WORDS:** Non-probability sample; Regression; Simulation; Small area estimation.

### 1. INTRODUCTION

County estimates of farm production are more and more in demand by government agencies for use in local economic decision making and by companies selling fertilizers, pesticides, crop insurance, and farm equipment. The United States Department of Agriculture (USDA) is currently implementing a program to standardize and improve county estimates of farm production (Bass *et al.* 1989). County estimation programs in the past have been carried out individually within each state. Because of this there has been little consistency across states in data collection and estimation methods used to produce county estimates. The goal of the USDA program for county estimation is to provide a set of sampling and estimation procedures for the states so that county estimation programs across the United States may yield estimates of comparable quality.

The new USDA county estimation program encompasses every stage of the production of county estimates from the construction of sampling frames through the estimation itself. The research described here is concerned only with the estimation of bushels of wheat produced. We hope, however, that our methods may prove useful in other aspects of the county estimation program, for example in estimating acres planted and for crops other than wheat.

Although the county estimation procedures used in the past varied from state to state, some parts of the procedures were similar. A typical procedure involved obtaining initial estimates from the data available within each county. Then an expert would review the estimates, alter them in light of his personal knowledge of the farms in the sample, weather conditions, and other factors, and then note the implications of the adjustments on the estimated total production for the state. The expert might repeat this process for a number of iterations until the

<sup>1</sup> Elizabeth A. Stasny, Prem K. Goel and Deborah J. Rumsey, Department of Statistics, The Ohio State University, 1958 Neil Avenue, Columbus, Ohio 43210, USA.



estimates within each county seemed reasonable and the resulting state production total agreed with the USDA state estimate. (The USDA state total is estimated based on a large probability sample and is thus thought to be a more accurate estimate than the total based on the county estimation procedure. For this reason, states typically constrain their county estimates to sum to the USDA estimate.)

Written documentation of the current county estimation procedures, as outlined above, typically is not available. Thus the assumptions and methods that the expert uses can not be inspected by others and it is practically impossible to study the procedures or replicate calculations. In addition, one cannot obtain variance estimates or use the procedures of one state in another state. New methods for county estimation must address these problems.

The data that we use in this research were collected in Kansas in 1987, before the new USDA county estimation sampling procedures were in use. Data from 1987 were used because the United States Agricultural Census was taken in that year and we may, therefore, use the Census data in our estimation procedure. Kansas data were chosen for use in this study because the county data collection program in Kansas was one of the more comprehensive programs in the United States. Nevertheless, the data used for county estimation in Kansas, as in most other states, were not collected from a probability sample of farms. Therefore, our estimation procedure must not require a probability sample of wheat farms. Such a procedure may also be useful under the new county estimation program since states still will not be required to choose probability samples of farms.

There is much recent research on small area estimation (see for example Platek *et al.* 1987). Standard small area estimation procedures, however, require known selection probabilities since the inverses of these probabilities are used to weight observations in standard estimators such as synthetic and direct estimators. (See for example Section 2 of Särndal and Hidiroglou 1989 for a discussion of standard small area estimators.)

The methods considered here must be different from the usual small area estimation techniques. First, the sample of farms available to produce county estimates is not typically a probability sample. Second, the county estimates must be constrained to sum to the USDA-produced state totals. Since most state agriculture departments currently do not have large computing facilities, an additional preliminary constraint on the estimation procedure is that computations must be simple enough to be performed on a personal computer. Thus, for our initial efforts, we prefer to avoid computationally intensive estimators such as those described by Fay and Herriot (1979). For these reasons, we consider a computationally simple estimator based on a regression model for producing county estimates of wheat production.

In Section 2 of this paper we describe the Kansas data bases used in this study. Section 3 presents the regression procedure for estimating wheat production while Section 4 describes several methods for scaling those estimates to the USDA state total. In Section 5 estimates from the regression models are obtained and compared to the published county estimates and to estimates produced using the synthetic estimator and the direct estimator. In Section 6 we present the results of a simulation study conducted to compare these same estimators. Section 7 gives conclusions and areas for future research.

## 2. KANSAS DATA

For the purpose of reporting farm production, all states are divided into nine or ten districts. Kansas is divided into nine districts such that each of the 105 counties in Kansas is completely contained within one of the districts. The locations of the districts and the number of counties within each district are as shown below:

<u>District Number</u>	<u>District Location</u>	<u>Countries in District</u>
1	Northwest	8
2	West Central	9
3	Southwest	14
4	North Central	11
5	Central	11
6	South Central	13
7	Northeast	11
8	East Central	14
9	Southeast	14

Two data bases which are used in the production of Kansas county estimates, the Planted Acres Data Base and the Small Grain Data Base, were available for our use in this research. Most of our work was done with 1987 data but we also verified our results with the 1988 data. The 1987 Planted Acres Data Base contains information on planted acreage for 37,094 farms throughout Kansas. (A farm is defined by USDA to be any place with annual sales of agricultural products of \$1,000 or more.) Of these farms, the 22,300 that reported planting some wheat were used in the simulation study described in Section 5. The 1987 Small Grain Data Base contains production information for 5,802 farms which reported planting small grain crops. Of these, the 1,707 that reported planting some wheat were used in our study.

Records on the Planted Acres Data Base are a composite of Kansas farm data from a number of sources collected at a number of times. First a list of names and addresses of farms is created using data collected by county appraisers. This data may be replaced and/or corrected using data from the Quarterly Agricultural Surveys and from Monthly Farm Reports. The Quarterly Agricultural Surveys use stratified systematic samples of approximately 2,600 farms. The response rate is approximately 80%. The Monthly Farm Report is completed by about 3,000 farmers who have agreed to file the reports. The same farmer may complete monthly reports for many years. The most recent data for each item appears in the Planted Acres Data Base and the record for any one farm in any year may contain information from a number of sources.

The 1987 Small Grain Data Base contains information on acres planted, acres harvested, and bushels produced for farms responding to the Quarterly Agricultural Surveys and the Kansas Small Grain Survey. About 6,000 surveys were mailed to a random sample of farms for the 1987 Kansas Small Grain Survey; about 50% of the surveys were completed and returned.

In addition to the potential problem with nonresponse bias in the Small Grain Data Base, there is typically a problem with response bias. The production reported by farmers is often lower than the actual production. The non-standard sample, nonresponse bias, and response bias lead us to develop the county estimation procedure described in the following sections.

**3. REGRESSION MODELING**

We propose the development of a regression model for use in producing county estimates. The calculations for fitting a multiple regression model can be performed using a number of statistical packages available for personal computers. In addition, our proposed estimator allows for the fact that we do not have a probability sample of farms and will produce county estimates that sum to the desired state total.

The steps in our procedure are as follows:

- 1) Use multiple regression to model the relationship between farm production and some predictor variables using the non-probability sample of farms.
- 2) Assume that the regression relationship holds for the entire population of farms in the state, and estimate farm production for all farms in each county.
- 3) Adjust the estimates of farm production to sum to the USDA state total.

To describe the regression model we need the following notation. For  $i = 1, 2, \dots, I$  ( $I = 105$  counties in Kansas) and  $j = 1, 2, \dots, n_i$  let

$n_i$  = number of farms from  $i^{\text{th}}$  county in sample;

$$n = \sum_{i=1}^I n_i = \text{total sample size};$$

$N_i$  = total number of farms from  $i^{\text{th}}$  county in population;

$$N = \sum_{i=1}^I N_i = \text{total number of farms in population};$$

$Y_{ij}$  = wheat production of  $j^{\text{th}}$  farm in  $i^{\text{th}}$  county (in bushels);

$X_{ij} = (1 \ X_{ij1} \ X_{ij2} \ \dots \ X_{ijp}) = \text{vector of } p \text{ predictors for } j^{\text{th}} \text{ farm in } i^{\text{th}} \text{ county.}$

It is important, as we will see later, to choose predictor variables for which county totals are known or for which very accurate estimates of the county totals are available. The predictor variables must also include information related to the probability that a farm is included in the sample, such as a measure of the size of a farm. This will allow us to use the regression model to adjust for the fact that the sample is not a probability sample.

We consider regression models of the form

$$Y_{ij} = f(X_{ij} | \beta) + \epsilon_{ij},$$

where  $\beta = (\beta_0 \ \beta_1 \ \beta_2 \ \dots \ \beta_p)$  is a vector of parameters and  $\epsilon_{ij}$  is a random error term with variance  $\sigma^2$ . Let the fitted values, which will be obtained using data from the Small Grain Data Base, be denoted by

$$\hat{Y}_{ij} = f(X_{ij} | \hat{\beta}).$$

Then the county total for the  $i^{\text{th}}$  county may be estimated as follows:

$$\hat{Y}_{i+} = \sum_{j=1}^{N_i} \hat{Y}_{ij} = \sum_{j=1}^{N_i} f(X_{ij} | \hat{\beta}),$$

where a “+” in a subscript indicates summation over the corresponding subscript.



For a general form of  $f(X_{ij} | \beta)$ , it would be necessary to know the value of  $X_{ij}$  for all farms in the  $i^{\text{th}}$  county. It is, of course, not possible to have such extensive information. If, however,  $f(X_{ij} | \beta)$  is a linear function, then we only need to know county totals of the predictor variables. This is the case since, for a linear regression equation,

$$\begin{aligned}\hat{Y}_{i+} &= \sum_{j=1}^{N_i} \hat{Y}_{ij} = \sum_{j=1}^{N_i} [\hat{\beta}_0 + \hat{\beta}_1 X_{ij1} + \hat{\beta}_2 X_{ij2} + \dots + \hat{\beta}_p X_{ijp}] \\ &= \hat{\beta}_0 N_i + \hat{\beta}_1 X_{i+1} + \hat{\beta}_2 X_{i+2} + \dots + \hat{\beta}_p X_{i+p},\end{aligned}$$

where  $X_{i+k}$  is the total of the  $k^{\text{th}}$  predictor for the  $i^{\text{th}}$  county.

The  $\hat{Y}_{i+}$  will be reasonable county estimates if the regression model describes the relationship between the predictor variables and production for all farms in each county as well as for the farms in the data base. These county estimates, however, will not necessarily sum to the USDA state total for production. Methods for resolving this problem will be considered in Section 4.

In addition to providing county estimates of farm production, the linear regression model proposed above also permits us to obtain variance estimates. This is easiest to see if we write the county estimates in terms of matrices. Let

$X = n \times (p + 1)$  matrix of actual data with rows being the  $X_{ij}$  defined above;

$Z = (\text{unknown}) N \times (p + 1)$  matrix of predictor variables for all farms in the state;

$\hat{Y} = (\text{unknown}) N \times 1$  vector of estimates of wheat production for all  $N$  farms in state;

$B_i = N \times 1$  column vector with elements  $b_{ij}$

$$\text{where } b_{ij} = \begin{cases} 1 & \text{if the } j^{\text{th}} \text{ farm is in the } i^{\text{th}} \text{ county} \\ 0 & \text{otherwise} \end{cases}$$

$$A = [B_1 B_2 B_3 \dots B_I]_{N \times I}.$$

The estimation procedure described above does not provide  $\hat{Y}$  but instead provides a vector of county estimates  $\hat{Y}_{i+} = A^T \hat{Y}$ , where “ $T$ ” indicates the transpose of a matrix.

The variance for the county estimates is thus

$$\text{Var}(\hat{Y}_{i+}) = \text{Var}(A^T \hat{Y}) = A^T \text{Var}(\hat{Y}) A = A^T \text{Var}(Z\hat{\beta}) A = \sigma^2 A^T Z (X^T X)^{-1} Z^T A.$$

Although  $Z$  itself is unknown, the product  $A^T Z$  is a known matrix containing only the numbers of farms in a county,  $N_i$ , and the county totals,  $X_{i+k}$ , for the predictor variables. Thus, if we use the regression mean square error (mse) as an estimate of  $\sigma^2$ , we may obtain estimates of the variances of the county estimates. Variance estimates for county estimates have not previously been available.

The estimator based on a regression model as described in this section meets the requirements for a computationally simple estimator from a non-probability sample. In the following section we consider methods to adjust the estimates to sum to the USDA state totals for farm production.

#### 4. SCALING ESTIMATES TO SUM TO STATE TOTAL

Let  $Y$  be the USDA's estimated total wheat production for Kansas. In general,  $\sum_{i=1}^I \hat{Y}_{i+} \neq Y$ . Thus, we define new estimates

$$\tilde{Y}_{i+} = c_i \hat{Y}_{i+},$$

where the  $c_i$  are constants such that  $\sum_{i=1}^I \tilde{Y}_{i+} = \sum_{i=1}^I c_i \hat{Y}_{i+} = Y$ . An important question is how to choose the  $c_i$ . Current methods used for county estimation take  $c_i = c$  (at the district level) and thus adjust all estimates by a common proportion. Instead, one could choose the  $c_i$  to minimize the sum of the squared differences or relative differences between the  $\tilde{Y}_{i+}$  and  $\hat{Y}_{i+}$ . Values of  $c_i$  and  $\tilde{Y}_{i+}$  for three criterion for choosing  $c_i$  are given below.

##### 1) Choose $c_i = c$

If  $c_i$  is taken to be a constant, then it is easy to show that

$$c_i = c = Y / \sum_{i=1}^I \hat{Y}_{i+}$$

and

$$\tilde{Y}_{i+} = Y \left( \hat{Y}_{i+} / \sum_{i=1}^I \hat{Y}_{i+} \right).$$

##### 2) Choose $c_i$ to minimize the sum of squared differences between $\tilde{Y}_{i+}$ and $\hat{Y}_{i+}$

To choose  $c_i$  to minimize the sum of the squared differences between  $\tilde{Y}_{i+}$  and  $\hat{Y}_{i+}$  subject to  $\sum_{i=1}^I c_i \hat{Y}_{i+} = Y$ , we must minimize  $\sum_{i=1}^I (\tilde{Y}_{i+} - \hat{Y}_{i+})^2 = \sum_{i=1}^I (c_i \hat{Y}_{i+} - \hat{Y}_{i+})^2$  with respect to  $c_i$  using a Lagrange multiplier to impose the desired constraint. Doing this, we find that

$$c_i = 1 + \left[ \left( Y - \sum_{i=1}^I \hat{Y}_{i+} \right) / \hat{Y}_{i+}^2 \sum_{i=1}^I (1 / \hat{Y}_{i+}) \right]$$

and

$$\tilde{Y}_{i+} = \hat{Y}_{i+} + \left[ \left( Y - \sum_{i=1}^I \hat{Y}_{i+} \right) / \hat{Y}_{i+} \sum_{i=1}^I (1 / \hat{Y}_{i+}) \right] \hat{Y}_{i+}.$$

Note that the scaled estimates,  $\tilde{Y}_{i+}$ , are obtained by adjusting the original estimates,  $\hat{Y}_{i+}$ , by adding a factor which is a proportion of the difference between the USDA state total and the sum of the original county estimates. The proportion is based on the harmonic mean of

the original estimates. Although some of these scaled estimates could be negative in theory, this is not considered likely in practice because farmers often underreport the amount of production on their farms. If the total of the original estimates exceeds the USDA state total, then scaled estimates corresponding to counties with small original estimates may be negative.

### 3) Choose $c_i$ to minimize the sum of squared relative differences between $\tilde{Y}_{i+}$ and $\hat{Y}_{i+}$

To choose  $c_i$  to minimize the sum of the squared relative differences between  $\tilde{Y}_{i+}$  and  $\hat{Y}_{i+}$  subject to  $\sum_{i=1}^I c_i \hat{Y}_{i+} = Y$ , we must minimize  $\sum_{i=1}^I [(\tilde{Y}_{i+} - \hat{Y}_{i+})/\hat{Y}_{i+}]^2 = \sum_{i=1}^I (c_i - 1)^2$  with respect to  $c_i$  using a Lagrange multiplier to impose the desired constraint. Doing this, we find that

$$c_i = 1 + \left[ \hat{Y}_{i+} \left( Y - \sum_{i=1}^I \hat{Y}_{i+} \right) / \sum_{i=1}^I \hat{Y}_{i+}^2 \right]$$

and

$$\tilde{Y}_{i+} = \hat{Y}_{i+} + \left[ \hat{Y}_{i+}^2 \left( Y - \sum_{i=1}^I \hat{Y}_{i+} \right) / \sum_{i=1}^I \hat{Y}_{i+}^2 \right]$$

The scaled estimates,  $\tilde{Y}_{i+}$ , are again obtained by adjusting the original estimates,  $\hat{Y}_{i+}$ , by adding a factor which is a proportion of the difference between the USDA state total and the sum of the original county estimates. The proportion here is based on the squared values of the original estimates. As in method 2, these scaled estimates may be negative, although it is unlikely in practice.

Note that we have chosen to consider the difference  $\tilde{Y}_{i+} - \hat{Y}_{i+}$  relative to  $\hat{Y}_{i+}$  rather than to  $\tilde{Y}_{i+}$ . This choice was made because in the later case the estimator,  $\tilde{Y}_{i+}$ , does not have a closed-form solution. Thus, to meet the goal of developing computationally simple estimators, we chose to consider the difference  $\tilde{Y}_{i+} - \hat{Y}_{i+}$  relative to  $\hat{Y}_{i+}$ .

In the following section we will consider the effects of these three scaling methods on the county estimates of wheat production.

## 5. COMPARISON OF ESTIMATES OF WHEAT PRODUCTION

We used a linear regression model, as described in Section 3, to model the relationship between wheat production (measured in total bushels produced) and some predictor variables for farms in the 1987 Small Grain Data Base. The possible predictor variables that we considered included: acres planted in wheat, acres of wheat harvested, a prediction of wheat production based on the 1986 county estimates, acres of irrigated wheat, acres of non-irrigated wheat, indicators of the district in which the farm is located, indicators of region of the state (east, central, west), and interaction terms.

The most important predictor variables for the regression model were acres planted in wheat and some indicator of the location of the farm within the state. The variable based on the previous year's county estimates did not seem to be a useful predictor for the amount of wheat produced on a farm in the current year. Because other possible predictor variables, such as irrigated acres, are not known as accurately at the county level, we decided that acres planted would be the single continuous predictor variable included in the model. Not all district indicators were needed in the regression model; that is, some districts were similar and could



**Table 1**  
Regression Models Fitted to Actual Data

	Fitted Models	$R^2$	$\sqrt{\text{mse}}$
Model 1	$\begin{aligned} \text{Bushels} = & -811 + 32(\text{Pla}) + 3,248I_1 + 3,088I_2 + 2,190I_3 \\ & + 2,526I_4 + 1,241I_5 - 562I_6 + 1,047I_7 \\ & + 399I_8 \end{aligned}$	85	5,945
Model 2	$\begin{aligned} \text{Bushels} = & -281 + 28(\text{Pla}) + 138I_1 + 1,861I_2 + 2,328I_3 \\ & + 329I_4 - 359I_5 - 334I_6 - 42I_7 + 500I_8 \\ & + 11(\text{Pla})I_1 + 5(\text{Pla})I_2 + 3(\text{Pla})I_3 + 11(\text{Pla})I_4 \\ & + 9(\text{Pla})I_5 - 0.2(\text{Pla})I_6 + 15(\text{Pla})I_7 - 7(\text{Pla})I_8 \end{aligned}$	86	5,818

Note: Pla is planted acres,  $I_i$  is the indicator variable for the  $i^{\text{th}}$  district.

have been grouped together. We decided, however, to include all district indicators in the model since groupings of districts might change from year to year or might be different for crops other than wheat.

We chose to focus our study on two possible regression models: Model 1 contained acres planted in wheat and the district indicators while Model 2 contained these same variables and the interaction terms involving acres planted and the indicator variables. The models and measures of their fits are shown in Table 1. Although the root mean squared errors did not differ considerably for the two models, we felt that the difference might be magnified when the models were used to estimate farm production for the entire state. Thus, in the following, we obtain and compare estimates from both models.

To verify that these regression models are not simply a result of some unusual feature in the 1987 Kansas Small Grain Data Base, we used the same set of possible predictor variables and searched for reasonable regression models using the 1988 data. The fits of Models 1 and 2 to the 1988 data are similar to the 1987 fits and no other model appeared to be superior for fitting the 1988 data. The estimates for the parameter corresponding to acres of wheat planted were fairly similar in both 1987 and 1988, but the parameters corresponding to the indicator variables for districts showed considerable change. We believe that the indicator variables for districts are reflecting the effects of weather and different farming practices in different parts of the state. For example, irrigation is more commonly used in western and central Kansas than in eastern Kansas. Although farming practices are not likely to change dramatically from one year to the next, weather conditions may be quite different. Thus, it seems reasonable that the contribution of the district variable in predicting wheat production could change considerably from year to year.

Both models were used to obtain county estimates for all 105 counties in Kansas. In Table 2, the unscaled estimates and their standard errors under both Models 1 and 2 are given for nine counties, one county chosen at random from within each district so that the nine counties are spread over the entire state. An inspection of Table 2 suggests that the estimated standard error for Shawnee county is an anomaly. The variance of a county estimate depends on the number of farms in the county, the total acres planted in wheat in the county, and the number of farms sampled from the district in which the county lies. District 8, in which Shawnee county is located, had relatively few farms in the Small Grain Data Base. The county has a moderate number of farms growing wheat but these farms are small in terms of acres planted. These three factors together result in the rather large standard error for the estimates from Shawnee county.

**Table 2**  
**Regression Model Estimates for Nine Counties in Kansas**

District	County	Estimated Bushels of Wheat Produced (in thousands of bushels)	
		Model 1 (no interaction terms)	Model 2 (with interaction terms)
1	Decatur	4,944 (180)	4,778 (179)
2	Trego	4,378 (174)	4,229 (188)
3	Hodgeman	4,808 (123)	4,908 (125)
4	Jewell	5,555 (275)	5,550 (269)
5	Marion	5,144 (313)	4,931 (315)
6	Comanche	2,615 (59)	2,480 (63)
7	Leavenworth	231 (53)	262 (61)
8	Shawnee	232 (106)	226 (104)
9	Butler	2,374 (331)	2,272 (338)

**Note:** Standard errors are given in parentheses below each estimate.

The estimates shown in Table 2 are reasonably similar to the published county estimates (Kansas Agricultural Statistics 1988). While it is encouraging that our estimates are not wildly different from those published by Kansas, there is no theoretical basis for using the Kansas estimates as a standard. Thus, we carried out a simulation study to help us evaluate our estimators. This study is described in the following section.

**6. SIMULATION STUDY**

**6.1 The Estimators to be Compared**

In the simulation study, we compared the estimates from our two regression models with those from two standard small area estimators: the synthetic and direct estimators. (See, for example, Section 2 of Särndal and Hidiroglou (1989) for a discussion of standard small area estimators, including the synthetic and direct estimators.) The synthetic estimates are obtained by allocating the state total for wheat production to the counties according to the proportion of total acres planted in wheat within each county. The direct estimates are obtained using only the sampled farms in a county to estimate wheat production for that county.

We expect the synthetic estimates to have a large amount of bias because counties in different parts of the state have different farming practices and different weather conditions, while the synthetic estimator treats each county as if it were representative of the entire state. The synthetic estimates, however, will have relatively small variances because they are obtained using all the data from the entire state.

Since the direct estimate for a county is based only on the sample data within that county, it will have a relatively large variance but it should have smaller bias than the synthetic estimate. At least one farm from a county must appear in the sample to make it possible to obtain an estimate for that county, and at least two farms are needed in the sample to make variance estimation possible. In the 1987 Kansas Small Grain Data Base, three counties had no wheat farms in the sample and three additional counties had only a single farm in the sample. Although we are comparing our regression model estimates to the synthetic and direct estimates, it should be noted that the latter two estimators require that the data be from a probability sample. This requirement is not met by the Kansas data.

Table 3  
Numbers of Farms and Production Levels by District and Planted Acres

District		Planted Acres in Farm				
		0-99	100-249	250-499	500-999	≥ 1,000
1	$M_i^*$	354	638	531	302	85
	$m_i^*$	27	45	51	40	9
	bu/pa*	34.68	37.18	37.76	39.21	38.68
2	$M_i$	266	550	572	377	161
	$m_i$	27	49	47	55	33
	bu/pa	35.92	33.62	36.78	39.09	34.85
3	$M_i$	264	549	610	537	264
	$m_i$	31	80	76	98	61
	bu/pa	26.93	32.84	35.03	36.79	33.13
4	$M_i$	956	939	626	271	50
	$m_i$	62	37	23	21	7
	bu/pa	36.81	36.91	39.70	39.87	39.41
5	$M_i$	1,236	1,529	912	350	54
	$m_i$	92	93	51	26	3
	bu/pa	31.79	32.25	31.69	36.85	33.65
6	$M_i$	1,181	1,427	1,160	793	249
	$m_i$	96	96	81	55	20
	bu/pa	26.24	26.88	28.78	27.87	26.72
7	$M_i$	957	242	67	9	3
	$m_i$	62	5	2	0	0
	bu/pa	33.87	40.81**	40.81**	40.81**	40.81**
8	$M_i$	1,126	251	52	9	1
	$m_i$	56	11	2	0	0
	bu/pa	26.02	11.48**	11.48**	11.48**	11.48**
9	$M_i$	1,122	431	166	59	12
	$m_i$	47	19	7	3	1
	bu/pa	23.57	23.87	27.63**	27.63**	27.63**

\*  $M_i$  is the number of farms on the Planted Acres Data Base,  $m_i$  is the number of farms in the Production Data Base, and bu/pa is the ratio of bushels produced to acres planted.  
\*\* Cells of this district were grouped to obtain bu/pa values.



## 6.2 The Simulated Population and Samples

We first simulated a population of wheat farms by generating production values for all 22,300 farms reporting acres planted in wheat on the Planted Acres Data Base. Because production rates appear to vary by district and size of farm (see Table 3), we generated bushels-per-planted-acres (bu/pa) from 37 different distributions. These distributions were based on the bu/pa data from the Small Grain Data Base. (Notice that in the eastern districts of Kansas, districts 7, 8 and 9, there were few or no sampled farms in several size-of-farm classifications. Those classifications were grouped as indicated in Table 3 for the purpose of simulating bu/pa values.) Histograms of the observed bu/pa from the Small Grain Data Base were generated by district and five sizes of farm: 0-99, 100-249, 250-499, 500-999, and 1000 or more acres of wheat planted. Since these histograms generally appeared mound-shaped, we chose to use normal distributions to model the distributions of the bu/pa. The means and variances of the normal distributions were taken to be the sample means and variances of bu/pa from the wheat farms in the Small Grain Data Base within the 37 district by size-of-farm classifications.

After the bu/pa values were generated from the appropriate normal distributions for each farm, the bushels of wheat produced were obtained by multiplying the simulated bu/pa by the reported acres planted in wheat for each farm. Ten samples were generated from the resulting simulated population. Since there was no sampling design to follow in creating these samples, we sampled each farm within the district by size-of-farm classifications with probabilities equal to the observed frequencies with which farms on the Planted Acres Data Base appeared in the Small Grain Data Base. That is, farms within classification *C*, say, were chosen to be in the sample with probability equal to

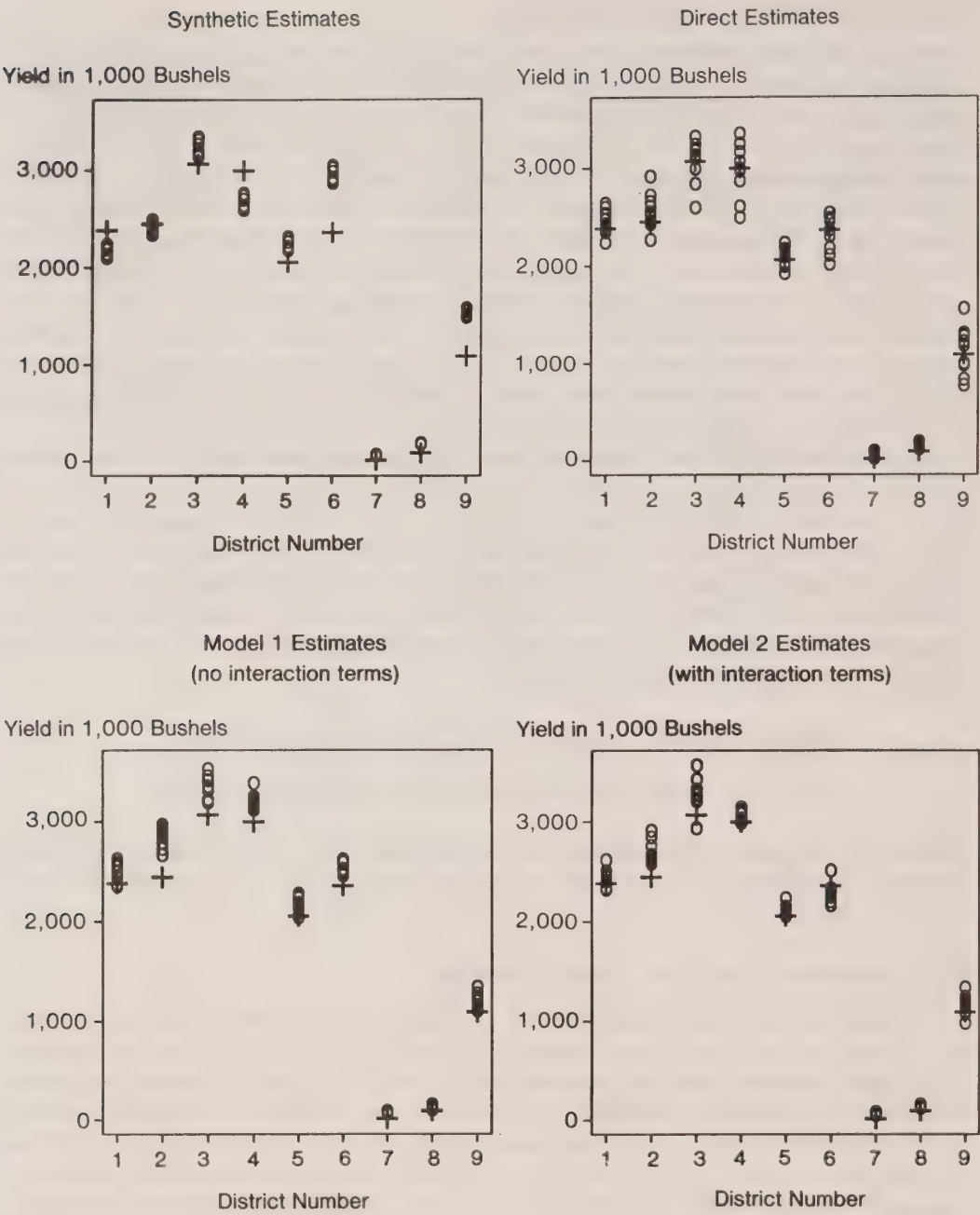
$$\frac{\text{Number of farms in classification } C \text{ in Small Grain Data Base}}{\text{Number of farms in classification } C \text{ in Planted Acres Data Base}}.$$

Our goal in using such a sampling scheme was to make the simulated samples as similar as possible to actual samples even though we do not know what the selection probabilities for the actual samples were.

## 6.3 Comparison of the Four County Estimators

We used the four county estimators (the two regression, the synthetic, and the direct) to obtain wheat production estimates for all 105 counties from each of the ten simulated samples. The resulting estimates were then compared to the "true" production values obtained for each county from the simulated population. This comparison allows us to evaluate the amounts of bias and variability in the estimates for each county. Figure 1 presents the values of all four estimates from each of the ten samples along with the true production values for the nine-randomly chosen counties, one from each district, which were previously mentioned in Section 5.

As expected, the synthetic estimates exhibit considerable bias. Indeed, only in district 2 does the range of estimates include the true population value. The ranges of the direct estimates are all larger than those of the synthetic estimates but those ranges do include the population values. The ranges of estimates from the regression models appear to be less than those of the direct estimates. For about half of the counties pictured in Figure 1, the estimates from Model 1 appear to exhibit some bias. The estimates under Model 2 seem to exhibit less bias. On the basis of this comparison of estimators we prefer Model 2, the regression model with the interaction terms.



**Note:** Estimates are for one county chosen at random from within each district.  
o = estimate from one of the ten simulated samples,  
+ = true value from simulated population.

**Figure 1.** Comparison of Estimators for Nine Counties

### 6.4 Comparison of the Scaling Methods

The same four sets of estimates for all counties from the ten sets of simulated samples were next scaled to agree with the state total from the simulated population using the three scaling methods described in Section 4. The resulting scaled estimates were compared to the true county production values for the simulated population. The comparison was made using the mean of the absolute value of relative error which is defined as follows:

$$(1/T) \sum_{i=1}^T \left| (\tilde{Y}_{i+} - Y_{i+}) / Y_{i+} \right|.$$

Figure 2 shows the values for all ten samples of the mean over the 105 counties of absolute relative error. This error is given for all four estimators under no scaling and under each of the three methods of scaling.

From Figure 2A, we see that the scaling method which minimizes the sum of squared differences produces very poor final estimates; the average of absolute relative differences between the final estimates and the county production values for the simulated population is quite large compared to that of the other scaling methods. This large error results from the fact that the total wheat production in one county may be quite different from that in another county. Since the scaling procedure minimizes the squared differences between the original and the final estimates, a county with a very small original estimate may have a final estimate that is changed considerably relative to the original estimate. These large changes in estimates do not seem warranted; hence we drop this method of scaling from consideration.

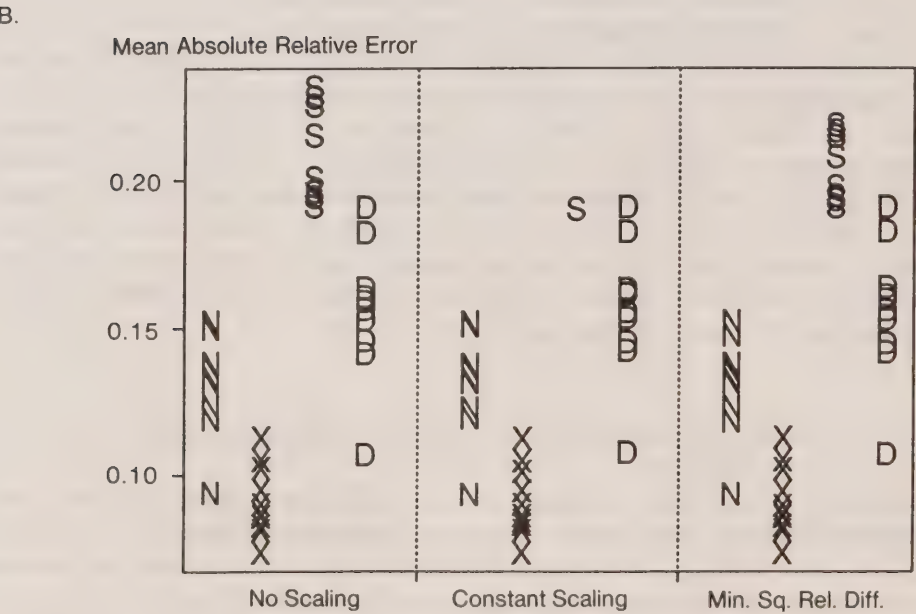
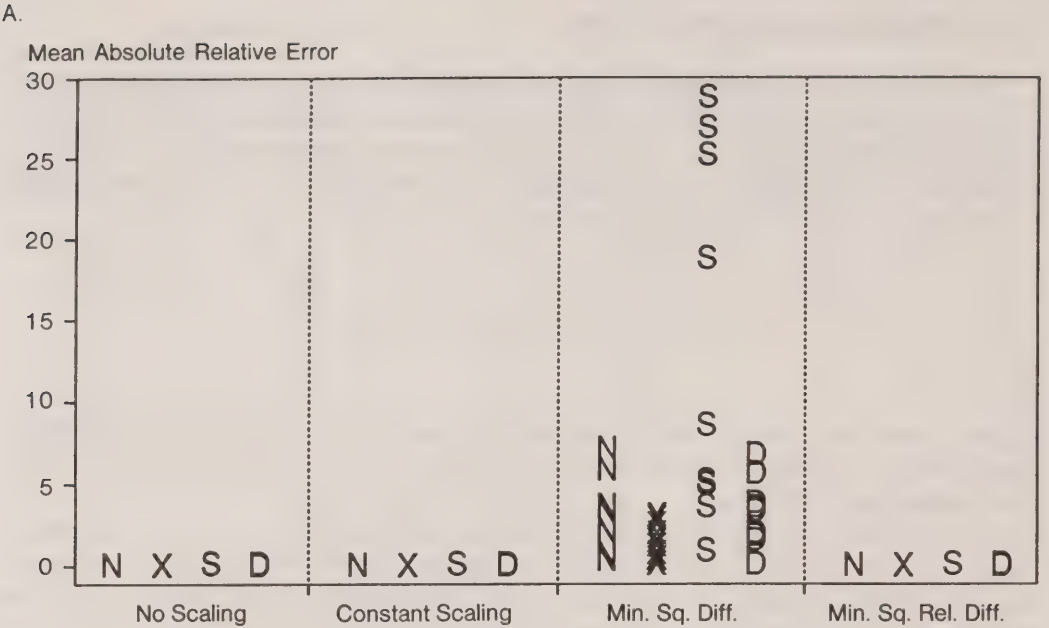
Figure 2B, a refinement of Figure 2A, provides a more detailed comparison of the four estimators under no scaling and under the two remaining scaling methods. We see from this figure that the error is generally smallest for the regression model with the interaction terms. This supports our choice of Model 2 in the previous subsection. In addition, Figure 2B suggests that there is little difference between the original unscaled estimates and the final estimates under either scaling method. In fact, the total of the original county estimates is not far from the simulated population total. Thus, the scaling constants,  $c_i$ , are all quite close to one. Since the two methods of scaling produce similar estimates, there is no reason to use the more difficult scaling method; the constant scaling method may be used.

## 7. CONCLUSIONS AND AREAS FOR FUTURE RESEARCH

We have shown that a regression model may be used to obtain reasonable county estimates of wheat production. The model we selected used acres planted, district indicators, and interaction terms as predictor variables. The regression model does not require a probability sample of farms and it does permit the estimation of variances of the county estimates. The estimates based on the regression model may be scaled to agree with state total production using a constant scaling factor since the alternative scaling method did not produce markedly different county estimates.

Many areas for future research in county estimation of farm production remain. For example, the county estimates from our simulation study suggested that the inclusion or exclusion of large farms (1,000 or more acres of wheat planted) from the sample for a district could have a large effect on the estimates for counties in that district. This was particularly true for





**Note:** Data are from the simulated samples.  
N = Model 1 Estimator (no interaction terms),  
X = Model 2 Estimator (with interaction terms),  
S = Synthetic Estimator,  
D = Direct Estimator.

**Figure 2.** Comparison of Estimators and Scaling Methods

districts which had few of these larger farms. Since large farms most likely account for a sizable proportion of farm production, it might be worthwhile to handle large farms separately in a county estimation procedure. States might also consider altering their sampling plans so that the largest farms are included in the samples with certainty.

Additional work is needed to determine whether a regression model similar to that developed for wheat is appropriate for other crops as well. In particular, it would be useful to discover if such models can be used for rare crops where there is much less available data. We should also note that the similarity in the state total and the total of the county estimates, which was observed for the actual data as well as for the simulated samples, may be characteristic of wheat production but not of all crop production. Future research should consider whether other crops require a scaling method other than constant scaling.

We chose to begin our research on the county estimation problem by studying methods of estimating production. An additional problem for future research is the estimation of total acreage planted for various crops. In this research we used 1987 agricultural census data to provide the needed information on numbers of farms and acres planted in wheat within each county. The agricultural census, however, is taken only every five years. In the intermediate years, changes in numbers of farms and acres planted must be estimated from sample data. We expect such changes in census values to be small for major crops like wheat in Kansas, but we anticipate greater difficulty estimating these quantities for less common crops.

Finally, the requirement for a computationally simple estimator, which led us to propose an estimator based on a regression model, may no longer be necessary as state agricultural offices are being linked to a large, national computer system. Thus, in our future research on county estimates of farm production, we plan to consider more computationally intensive small-area estimators.

### ACKNOWLEDGEMENTS

This research was supported in part by the United States Department of Agriculture under Cooperative Agreement No. 58-3AEU-9-80040. The authors take sole responsibility for the contents of this paper. The authors wish to thank Gary Keough and Leland Brown at USDA and Ronald Sadler, Melvin Perrott, Eldon Thiessen, and M. E. Johnson at the Kansas Department of Agriculture for their help on this project. We also thank the referee and associate editor for their helpful comments on an earlier version of this paper.

### REFERENCES

- BASS, J., GUINN, B., KLUGH, B., RUCKMAN, C., THORSON, J., and WALDROP, J. (1989). Report of the Task Group for Review and Recommendations on County Estimates. USDA National Agricultural Statistics Service, Washington, D.C.
- FAY, R.E., and HERRIOT, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.
- KANSAS AGRICULTURAL STATISTICS. (1988). Kansas Farm Facts, prepared by the Statistical Division of the Kansas Department of Agriculture in cooperation with the National Agricultural Statistics Service of the U. S. Department of Agriculture, Topeka, Kansas.
- PLATEK, R., RAO, J.N.K., SÄRNDAL, C.E., and SINGH M.P. (Eds.) (1987). *Small Area Statistics*. New York: John Wiley & Sons.
- SÄRNDAL, C.E., and HIDIROGLOU, M.A. (1989). Small domain estimation: A conditional analysis. *Journal of the American Statistical Association*, 84, 266-275.





## **Canada's General Social Survey: Five Years of Experience**

**D.A. NORRIS and D.G. PATON<sup>1</sup>**

### **ABSTRACT**

The Canadian General Social Survey is an annual survey that aims to provide data on the demographic and social characteristics of Canadians. This paper provides an overview of the program, based on the experience of the first five surveys. The objectives of the program, the methodology used, the themes and issues addressed, the program outputs and the plans for the future are all discussed.

**KEY WORDS:** Social surveys; Telephone surveys; Random digit dialing; Time use surveys; Health surveys.

### **1. INTRODUCTION**

Statistics Canada's social statistics program is concerned with providing information on the demographic and social characteristics and conditions of Canadians. The program's output sustains the development of policy on many critical social issues.

The Census of Population, held every five years, is the cornerstone of the social statistics program, providing benchmark information on the demographic, social, and economic conditions of the population and the basis for future sample surveys of the population. In addition to the Census, activities include on-going surveys and other statistical programs, many based on administrative data sources, in the areas of Health, Education, Culture, Justice, Public Finance, Employment and Unemployment, Income and Expenditures and Demography.

While household surveys have long been an important part of the social statistics program, the regular survey program has historically been directed mainly at labour market and income related issues and there have been no regular ongoing surveys in areas such as health, education, justice or culture. In order to partially fill this data gap Statistics Canada established in 1985 a General Social Survey (GSS) program.

The purpose of this paper is to outline the nature and scope of the GSS program and to describe its evolution over the past five years. Included is a description of the methodology and the content of the five surveys that make up the program. Finally there is a brief discussion of some future directions for the program.

### **2. GSS PROGRAM OBJECTIVES AND STRUCTURE**

The period 1930-1980 witnessed a rapid rise in the number and size of social programs in Canada. Whereas in the early 1930's all government expenditures on social programs accounted for about 10% of GNP, by the early 1980's this expenditure had climbed to about 30%. Along with this rise came an increased demand for and use of data and information to monitor and analyze social trends, and over the years, Statistics Canada expanded its social statistics program to meet growing requirements. Nonetheless, the more extensive use of available data in recent

---

<sup>1</sup> D.A. Norris and D.G. Paton, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6.

years revealed major areas of weakness where relevant data were too narrow and restrictive for the effective planning of policy programs, products, and services, or for determining the allocation of resources between competing alternatives.

In the early 1980's, a shortcoming of the social statistics program was that aside from the labour market and income areas, most other social data were derived from administrative records or surveys of institutions. These data sources provided only limited information on the population who came in contact with social institutions and no data on the need for, or impact of, social programs on the general population. Such data can only be obtained through a general population survey.

While a case could have been made for regular, frequent and large scale surveys in a variety of fields (*eg.* health, education, criminal victimization), resources to mount such a large scale program were not available. Instead Statistics Canada initiated a much more modest annual General Social Survey which over five years would cover major topics of importance and which would in the long term serve as a vehicle for monitoring social change. In the short term it could also serve as a vehicle to collect limited data on topics of current social policy interest. The total annual budget for the GSS was originally set at about one million dollars (CAN) and the program was funded by an internal reallocation of Statistics Canada resources derived from efficiency gains in the Labour Force Survey program.

The objectives of the GSS program are two-fold:

- To gather data with a degree of regularity on a broad range of social trends in order to monitor temporal changes in the living conditions and well-being of Canadians; and,
- To provide information on specific social policy issues of current or emerging interest.

To meet these objectives, the GSS program was established with an annual survey cycle. In order to cover the wide range of social issues for which data are required, the GSS program consists of five survey cycles, each covering a different core topic. The collection of data for these topics is thus repeated every five years. The core topics identified for the five cycles are:

1. Health
2. Time Use
3. Personal Risk (accidents and criminal victimizations)
4. Education and Work
5. Family and Friends.

An additional objective in planning content was to include questions that would be useful in deriving indicators of the quality of life, for example, measures of life satisfaction, attitudes, perceptions, or beliefs.

The content of a GSS cycle consists of the following three modules:

- Core content, which is repeated every five years in order to gather information to monitor trends in living conditions and well-being.
- Focus content, which varies from survey to survey and is aimed at the second survey objective of providing information on specific policy issues of particular interest to certain federal departments or policy groups.
- Classification content, which is collected in every cycle and consists of a set of basic demographic and socio-economic variables that enable the delineation of various population groups to facilitate the analysis of core and focus content.

While core and classification content are funded by Statistics Canada, costs associated with focus content are recovered from sponsors.

The target population for the GSS consists of the non-institutionalized population aged 15 and over living in the ten provinces. It was decided that the Labour Force Survey would not

be used as a vehicle for the GSS in order to avoid placing an excessive response burden on LFS respondents and to allow the GSS to use sampling and collection methodologies and sample allocations that differ from those of the LFS. The target sample size for each cycle is 10,000 individuals which was arrived at as a compromise between the competing demands of precision of estimates, budget and length of interview. However, there exists within the GSS program the potential for sponsors to expand the sample for a target population or geographic area. The first survey on health was conducted in late 1985 and the other surveys followed at approximately one year intervals. The fifth cycle on the family was conducted in early 1990 and data collection for Cycle 6 began in January 1991.

The themes and research issues which are covered by each of the surveys are discussed in more detail below. However, before considering these the methodology of the survey is examined in more detail.

### **3. METHODOLOGY**

#### **3.1 Requirements and Constraints**

The following are the principal methodological requirements of the GSS: i) it should allow for extensive analysis of the adult Canadian population at a national level and somewhat less detailed analysis at the regional level (this has implications on both the sample size and on the amount of data collected from each respondent); ii) it should have an acceptable cost; iii) it should have a design appropriate for a multipurpose survey; and iv) it should provide public use microdata sets that could be used for analysis by researchers outside Statistics Canada without too much difficulty.

These requirements all interact with the choice of data collection mode, sample design and sample size, but the last two were principally responsible for the choice of sample design, while the sample design and the first two requirements were largely responsible for the choice of data collection mode and sample size.

The last requirement suggests that the sample design be simple as the design information that would be necessary to analyse complex survey data cannot generally be made available on public use files. Requirement iii) suggests that the design not be highly optimized for specific variables.

#### **3.2 Mode of Data Collection**

The choice of data collection mode involved balancing a number of competing factors: cost per interview, length of interview, response rate, accuracy of information collected and sample size. The level of detail required in the data collected meant that interviews were expected to last 20 to 30 minutes per respondent. To reduce response burden at the household level and to avoid a cluster effect at the household level it was decided that only one person per household would be interviewed. The principal data collection methods considered for the survey were: self-completed mail-back questionnaire; personal interview; and telephone interview. The high non-response rates experienced with self-completed mail-back questionnaires were felt to be unacceptable (in terms of potential biases) given the heterogeneity of the target population. Personal interviews were felt to offer a number of advantages that would improve the quality of the data collected such as low non-response rates and low item non-response rates, but suffered from the disadvantage of high cost. In addition, many designs used to reduce the cost of personal interviewing have multiple stages of selection and are highly optimized for a few variables. (To not use a design and frame currently used for personal interviewing would have



been unreasonably costly.) These complicated designs make analysis of the resulting datasets difficult and the optimization leads to high design effects for some variables. These high design effects make such designs less appropriate for multipurpose surveys like the GSS. Experience with telephone surveys at Statistics Canada indicated that fairly high response rates could be achieved at reasonable cost. In addition, random digit dialing (RDD) sampling methods allow the efficient selection of samples that are simple random samples or nearly simple random samples.

For these reasons, the GSS has used telephone sampling (RDD) methods and telephone interviewing for most of its sample in all cycles conducted to date. When there has been a need to focus on special target groups its main sample has been supplemented with individuals selected from list frames. In Cycle 1 it was felt that face to face interviews should be used for many of the interviews with elderly respondents.

### 3.3 Target Population

The target population of the GSS is all persons over the age of 14 permanently living in Canada, with the following two exclusions: i) residents of the Yukon and Northwest Territories, and ii) residents of institutions. This target population is different from that of the Labour Force Survey, which in addition excludes residents of Indian Reserves and full-time members of the Canadian Armed Forces.

### 3.4 Sampled Population

The sampling methods used for the GSS exclude some members of the target population from the sample. During weighting, these exclusions are implicitly assumed to be similar to the sampled population (missing at random) and the final weights produce estimates for the target population.

When telephone interviewing methods are used, those persons living in households without telephones are excluded from the sample. This affects less than 2% of Canadian households covered by the Labour Force Survey (Statistics Canada 1989, 1990b). This high rate of telephone penetration is not uniform across age and income groups and varies from province to province: 95.4% of households in the province of Prince Edward Island have telephones while 99.2% of those in the province of Ontario do; 99.1% of households with incomes between 20 and 25 thousand dollars have telephones while only 93.9% of those with incomes less than 10 thousand do. Some subpopulations have much lower rates of telephone ownership than the average; for instance, only 86.7% of low income persons under 65 living alone have telephones.

The GSS does not in general accept proxy responses and so individuals who cannot use a telephone (those unable to hear or unable to speak) or who cannot be reached by phone during the survey period or who do not speak either English or French are excluded from the responding population. (For the sixth GSS cycle (on the health of Canadians) it was decided to accept proxy responses in those situations where the selected respondent could not complete the interview due to a health problem.)

When supplementary samples are drawn from lists of households interviewed by the Labour Force Survey (as was done for GSS Cycles 1, 5 and 6), residents of Indian Reserves and full-time members of the Canadian Armed Forces are excluded from these samples. These exclusions represent less than 0.5% of the population over the age of 65. (This is the only age group that has been sampled this way.)

### 3.5 Stratification

The stratification used by the GSS is determined by estimation requirements, operational requirements, restrictions on the definition of strata imposed by RDD sampling, weighting problems specific to RDD sampling and the special needs of sponsors. Since some estimates are required at the provincial level the GSS strata never cross provincial boundaries. For operational reasons a stratum must be interviewed from a single Regional Office, thus strata never cross Regional Office boundaries. The RDD sampling method used requires that strata be defined as aggregations of telephone exchanges. During weighting, accurate estimates of the sizes of strata are needed, thus the strata (defined on the basis of telephone geography) need to correspond closely to aggregations of units for which accurate population data or estimates were available. These accurate data are available in intercensal years at the Census Metropolitan Area (CMA) level.

The basic stratification based on these requirements starts with the provincial boundaries as stratum boundaries. In Cycles 1 to 5, Saskatchewan and Ontario were each covered by two Regional Offices, so they were both divided in two by a stratum boundary. Further, within each of the areas thus obtained, the CMA's formed a stratum and the non-CMA areas another stratum. In addition, the two largest CMA's, Montreal and Toronto, were each separate strata. For Cycles 1 to 5 this gives us a total of 25 strata: one in Prince Edward Island (there is no CMA in PEI), two in each of Newfoundland, Nova Scotia, New Brunswick, Manitoba, Alberta and British Columbia, three in Quebec, four in Saskatchewan and five in Ontario. For Cycle 6 there were 21 strata: one in Prince Edward Island, two in each of Newfoundland, Nova Scotia, New Brunswick, Manitoba, Saskatchewan, Alberta and British Columbia, and three in Quebec and Ontario.

This is the basic stratification used by the GSS, but modifications to this basic common stratification to accommodate the particular needs of the subject matter or of the sponsors are possible and have been implemented. In Cycle 2 the special interest in language use indicated that separate strata with higher sampling fractions should be used in "contact regions" in which there were thought to be large numbers of both anglophones and francophones. In Cycle 5 the interest of a client in producing estimates for certain sub-provincial regions of Ontario led to the definition of a special stratification.

### 3.6 Allocation

The target sample size of the GSS is 10,000 completed interviews. This sample has been allocated to provinces in proportion to the square roots of their population sizes. The allocation to strata within provinces has been in proportion to their sizes. The square root allocation is a method of increasing the sample sizes for the smaller provinces (when compared with a proportional allocation) without compromising the precision of Canada level estimates as much as an equal allocation. The method of Kish (1976) for arriving at an allocation that explicitly balances the need for provincial and Canada level precision has been investigated, but the resulting allocations yield little improvement in the precision at the Canada level while changing the allocations to some provinces dramatically and in a way felt to be undesirable.

### 3.7 Telephone Sampling Method

Except for supplemental samples of the population over 65 selected using lists of households interviewed for the Labour Force Survey, the GSS samples have been selected using random digit dialing methods. Two methods of sample selection have been used, the Waksberg (1978) method and the elimination of non-working banks method. Both methods use information obtained from telephone companies to improve the success rate of reaching households. The choice of methods depends on the level of detail of the information available.



Telephone numbers in Canada are ten digit numbers that can be decomposed into a three digit “Area Code”, a three digit “Prefix”, and two two digit fields, the first of which we refer to as a bank identifier. Thus within each “Area Code-Prefix” (ACP) there are ten thousand possible numbers and within each “Area Code-Prefix-Bank” (ACP-Bank or simply bank) there are one hundred possible numbers. For example, here is a fictitious telephone number and its components:

216-357-4675

216	Area Code
357	Prefix (exchange)
46	Bank Identifier
75	Number
216-357	ACP
216-357-46	ACP-Bank (bank).

When the only information that is available is a list of ACP’s, the GSS uses the Waksberg method of generating the sample. In this method, banks are selected with probability proportional to size, where the size measure is the number of residential telephone numbers in the bank. Within each selected bank a simple random sample of residential numbers is selected. When the sample size is the same in each bank, this method yields an equal probability sample of residential telephone numbers. The sample size within banks used by the GSS has been 6. This method has the advantage of improving the success rate of selecting residential numbers with the disadvantage of producing a clustered sample. For instance, in some rural areas of western Canada only approximately 6% of the numbers generated using lists of ACP’s are residential, while the success rate during the second stage of selection is about 50%. The design effects due to clustering are small for many variables, on the order of 1.0-1.3.

When more detailed information is available that allows the creation of a list of banks containing one or more residential numbers (“working banks”) the method which we call the elimination of non-working banks method (ENWB) is used. A simple random sample of numbers within the working banks is selected and non-residential numbers are rejected, yielding a simple random sample of residential numbers. Since the first GSS in 1985, sampling has shifted more and more to the ENWB method as more information has become available from the telephone companies. For Cycle 6 (conducted in 1991) the ENWB method was used for the entire sample.

A system of computer programmes for the Regional Offices of Statistics Canada has been written to implement these two sampling schemes and to monitor the progress of the survey. Within a stratum, the entire sample must be generated using the same sampling method.

After a household has been reached by telephone, a list of the names and ages of all household members is collected and, using this list and a set of random numbers printed for each questionnaire, one person 15 years of age or older in the household is selected to be interviewed. This is the method of Kish (1949).

3.8 Special Samples

Sponsors of the GSS have the opportunity to fund additional interviews. These additional samples can be simple increases in the RDD sample size for one or more strata or they can be drawn from other sampling frames.

In Cycles 2 and 5 the RDD samples in strata of special interest to sponsors were increased.

In Cycles 1, 5 and 6 additional samples of special interest groups were used to supplement the RDD sample. In these cases samples of persons aged 65 and over were selected using lists of households that had recently been part of the LFS sample.



### 3.9 Response Rates

One disadvantage of telephone surveys is that respondents seem to find it easier to refuse to participate in a telephone survey than in a survey with personal interviews. Telephone soliciting is being used regularly by businesses to sell products and services and everyone has to learn to say no over the phone. In addition, new technologies such as answering machines and special features being added to telephone systems are making it possible and easy for people to screen their incoming calls.

Table 1 gives response rates for the first five cycles of the GSS. The categories "Other Household Non-Response" and "Other Respondent Non-Response" include non-interviews due to language problems, illness, death in the family and absence for the survey period; some of these non-responses are undoubtedly refusals in disguise. In all cycles except Cycle 2 interviews were conducted as soon as possible after contacting the households. In Cycle 2 there was a gap of about a month between the initial contact with the households and the interviewing; there is a component of non-response that can be directly attributed to this time lag. From the table it seems that there may be a trend toward lower response rates over the five cycles.

If we consult Table 2, which presents response rates for individual Regional Offices for Cycles 3 to 6, we see that the situation is not so simple, with many offices (Halifax, Montreal, Winnipeg) showing little change in response rate over these cycles. In fact if we exclude the results obtained in Toronto, the response rate declined only slightly between Cycle 3 and Cycle 5. We have observed that more experienced interviewers tend to be more successful at achieving high response rates. The dramatic change in response rates over three cycles experienced by the Toronto office may in large part be due to the difficulty in hiring and retaining staff in a city that at the time had a booming economy. It is also possible that some of the change is due to a change in the population sampled from the Toronto office.

Preliminary results from eight (January to August 1991) months of data collection for Cycle 6 indicate (see Table 2) that it was possible to reverse the trend to lower response rates. There were a number of changes made between Cycles 5 and 6, the most important ones being a change to monthly data collection and the reassignment of the sample from offices not used for data collection in Cycle 6: St. John's sample was transferred to Halifax, Toronto's to Sturgeon Falls and Edmonton's to Winnipeg.

During data collection for Cycle 3 it was noted by interviewers that an increasing number of calls were answered by answering machines. This raised the concern that respondents might use these machines to screen their calls, resulting in higher non-response rates. We are not able

**Table 1**  
Response and Non-response Rates (%) by Cycle and Type

Result	Cycle				
	1	2	3	4	5
Household Refusal	6.2	6.2	6.0	7.2	10.3
Other Household Non-Response	4.4	6.8	6.6	6.4	7.2
Respondent Refusal	1.3	2.8	1.3	1.7	2.4
Other Respondent Non-Response	4.8	3.5	3.2	3.9	4.3
Special Cycle 2 Non-Response		1.9			
Response	83.4	78.9	82.9	80.7	75.8

**Table 2**  
**Response Rates by Cycle and Regional Office**  
 (results for Cycle 6 are preliminary – indicates offices not conducting interviews)

Regional Office	Cycle			
	3	4	5	6
St. John's	84.1	82.8	90.9	–
Halifax	84.7	84.1	85.9	82
Montreal	83.0	79.6	81.2	82
Sturgeon Falls	76.5	81.1	71.5	71
Toronto	87.0	75.4	63.0	–
Winnipeg	84.3	87.0	84.3	89
Edmonton	83.2	79.4	76.8	–
Vancouver	75.3	80.2	79.6	82
Canada	82.9	80.7	75.8	81
Canada (without Toronto)	82.1	81.8	80.1	–

**Table 3**  
**Response and Non-response Rates (%) by Type and Contact with Answering Machines**

	Did any calls reach an answering machine?			
	Cycle 4		Cycle 5	
	No	Yes	No	Yes
Household Refusal	7.18	8.17	10.34	9.74
Other Household Non-Response	6.45	5.89	7.19	7.41
Respondent Refusal	5.05	4.55	4.39	3.27
Other Respondent Non-Response	1.68	2.01	2.31	3.15
Responses	79.64	79.38	75.76	76.44
Number of Records	10,981 (93.6%)	747 (6.4%)	16,611 (90.6%)	1,715 (9.4%)

**Table 4**  
**Response and Non-response Rates (%) by Type and Type of First Contact**

	Was the first contact with an answering machine?			
	Cycle 4		Cycle 5	
	No	Yes	No	Yes
Household Refusal	7.24	7.19	10.46	7.90
Other Household Non-Response	6.46	5.40	7.15	8.06
Respondent Refusal	5.08	3.96	4.38	3.07
Other Respondent Non-Response	1.70	1.80	2.43	1.92
Responses	79.52	81.65	75.58	79.05
Number of Records	11,172 (95.3%)	556 (4.7%)	17,023 (92.9%)	1,303 (7.1%)

to identify those calls that were answered by a machine for cycles 1 to 3, but we are for subsequent cycles and so can analyze to some extent the effect of their use on response rates. Table 3 compares the response rates for those households for which none of the calls were answered by a machine with those for which at least one call was. No important effect of answering machines is indicated by this table; however the increase in contacts with answering machines, from 6.4% to 9.4% of households, is dramatic (Table 3). Table 4 compares the response rates for those households for which the first answered call was answered by a machine with those for which it was not. If any effect of answering machines is indicated by this table it is that response rates are higher for those households with a first contact by answering machine. There appears to be no evidence that the use of answering machines is seriously reducing response rates.

### 3.10 Data Capture and Processing

The data for all five cycles were captured directly into the mini-computers in Statistics Canada's regional offices. Some simple edits to check the validity of data as captured were made at the time of capture, but these could in most cases be overridden using special functions. Following transmission of the raw data to Ottawa an exhaustive set of edits was applied to find, and correct if possible, invalid or inconsistent responses. When a response was missing, invalid or inconsistent with other responses and the appropriate value could not be inferred from other responses on the questionnaire an 'unknown' code was assigned. Exceptions to this rule were three variables needed for weighting purposes: age, sex and number of telephone lines. In cases where these variables were missing the questionnaires themselves were consulted to assist in the imputation of values.

### 3.11 Weighting

#### 3.11.1 Initial Weights

Both the Waksberg and ENWB methods of selecting RDD samples yield self-weighting samples of residential telephone numbers. The Waksberg method does not provide an estimate of this weight, but for GSS weighting purposes it is sufficient to use an initial weight of one (1) for telephone numbers in those strata where that method is used. In ENWB strata the initial weight is the reciprocal of the probability of selection of the telephone number. This probability is simply:

$$\frac{n_c}{100 \times N_B},$$

where:

$n_c$  is the number of telephone numbers selected, and

$N_B$  is the number of working banks in the frame.

#### 3.11.2 Non-response Adjustment

The initial weight is adjusted for non-response using adjustment "strata" based on telephone geography. These are typically banks in Waksberg method strata and ACP's in ENWB strata. The initial weights are inflated by the following factor:

$$\frac{n_R + n_{NR}}{n_R},$$



where:

$n_R$  is the number of responding households in the non-response "stratum", and  
 $n_{NR}$  is the corresponding number of non-responding households.

### 3.11.3 Telephone Adjustment

Since households with more than one telephone line have a higher chance of being selected by an RDD survey, the initial weight adjusted for non-response (a weight for telephone numbers) is further adjusted by dividing by the number of telephone lines for the household to yield a household weight.

### 3.11.4 Initial Person Weight

Since only one eligible respondent per household is interviewed, the household weight must be adjusted by multiplying by the number of eligible respondents to yield a person weight.

### 3.11.5 Poststratification

At this point populations projected from the census are used as reference totals in the poststratification of the person weights, first to the stratum population sizes and then to the provincial age-sex populations. (It should be noted that it is only after the first stage of poststratification that the weights in Waksberg strata actually sum to a population estimate. Until this step they differ from a set of weights based on the inverses of the selection probabilities by an unknown constant of proportionality.) These two sets of reference totals are then used as the margins for a raking ratio adjustment to the weights.

## 4. THEMES AND RESEARCH ISSUES COVERED BY THE GSS

As indicated above, in order to cover a wide range of social issues, the GSS examines a different core topic each year for five years and then the topics are repeated. The core topics were chosen to fill perceived data gaps in the social statistics program. The five core themes are discussed in more detail below.

### 4.1 Health

The core content of the health cycle is directed at providing a range of measures of health status, including short and long term disability, the prevalence of common chronic conditions, such as high blood pressure or diabetes, and the use of various health care services. In addition, data are collected on life-style such as, smoking, drinking, and physical exercise. When linked to health status, these data provide information on the barriers (*e.g.* smoking, drinking) and bridges (*e.g.* physical exercise) to positive health for various population groups.

For the first GSS health cycle, the add-on focus content was directed at older Canadians and covered social networks, support given and received, as well as participation in a range of social activities. The sample size for the elderly population was also increased to allow for more in-depth analyses.

### 4.2 Time Use

The GSS time use survey consisted of a "24 hour time budget" generally for the day preceding the interview. Respondents provided information on each primary activity engaged

in during that day, the start time and duration of each activity, and associated information on where the activity took place and who was with the respondent at the time (*e.g.* spouse, children, friends, *etc.*). These data provide information on the frequency with which people participate in activities such as paid work, household work, attending cultural events, watching television, and the time spent on these activities.

The survey provides information on how Canadians allocate their time to activities such as paid work, housework and other non-market work and leisure activities. The data can be used to show constraints that limit a person's choice of the use of time and how these are distributed among different population groups. The inclusion of a battery of questions on satisfaction with various dimensions of life allows such measures to be correlated with patterns of time use for different population groups.

The 1986 GSS time use cycle also included a small module on intergenerational social mobility that allows for the analysis of movement on an occupational or educational hierarchy between the respondent and his or her parents.

The add-on focus content for the time use cycle was a detailed set of questions on language knowledge and use. While focus content is generally expected to be related to and complement core content, there was a demand for much more detailed language data than could be included in the population census. The information collected included data on language use at various stages of life (*e.g.* first learned, during childhood, at school) and in various settings including at home, at work, with friends, watching television, and in dealing with federal agencies. In order to allow a more detailed analysis in bilingual regions of the country, sample size was also increased in these geographic areas.

#### **4.3 Personal Risk**

The third GSS cycle was based around the topic of personal risk, including both criminal victimizations and accidents. Traditionally, information on these topics has been derived from administrative sources, such as police statistics and hospital records. However, these data provide very little information about the victim and, in addition, there are many crimes (the GSS estimates more than half) and accidents which are not reported to authorities.

The personal risk survey conducted in early 1988 asked respondents about criminal victimizations and accidents that they had experienced during calendar year 1987. Data were also collected on several life-style measures, such as alcohol consumption and frequency of night outings to allow these to be correlated with criminal victimizations and accidents. For each reported crime or accident incident, data were collected on the nature of the incident, the consequences in terms of activity restriction, medical attention and financial loss. In addition, respondents were asked to report their perceptions of crimes and accidents and about precautions taken to prevent these events.

The add-on focus content for the personal risk cycle was a set of questions on contact with the criminal justice system (*e.g.* police, courts, lawyers) and on the awareness and use of services by victims of crime.

#### **4.4 Education and Work**

While the monthly Labour Force Survey and other labour related surveys provide a wealth of information about the labour force, none of the existing surveys provides much information on the social aspects of work or the perceived quality of working life. The GSS cycle on education and work, conducted in early 1989, was designed to partially fill this data gap.

The survey was developed around three main themes that reflect fundamental changes in Canadian society: patterns and trends in work and education; new technologies and human resources; and work in the service economy. The themes reflect a range of issues on which more information is required. For example, the accelerating rate of technological innovation demands detailed knowledge about the utilization of and training for computers. Concerns about the effective utilization of the nation's human capital require a better understanding of the links between the labour force and the educational system. We also must anticipate future demands on educational institutions, and changing relationships between educational attainment and socio-economic outcomes. This round of the GSS also augments existing data sources by providing new information about the elderly population as well as some of the socio-economic implications of the baby boom generation entering middle age.

The survey collected a partial work and education history. It also included information on technology training and the use of computers, and on future plans for education. Subjective information also was sought in the form of a series of questions about satisfaction with retirement and other dimensions of life, as well as a block of questions on attitudes to science and technology.

#### **4.5 Family and Friends**

The fifth cycle of the GSS was based around issues related to family and friends and was completed in early 1990. While the Census and other household surveys provide family-based data, changes in family life have resulted in a need for new types of information. One short-coming of existing data is that generally they are based on a rather narrow concept of the family, in particular a nuclear family of parents and children or perhaps an economic family of related individuals living in the same household. This survey looks at the family in a broader context and collects information on the extent and nature of kinship networks and related questions of patterns of informal help and support among family and friends.

A second major theme of the survey is a result of the trends in marriage, divorce, and the increased frequency of common law unions. Increased numbers of Canadians are living in more than one union during their life time. The impact of such changes on family life and children is substantial and can best be studied by an analysis of marital and family history data. Such data were retrospectively collected in a special Family History Survey conducted in 1984 (Burch 1985). The GSS family cycle incorporates the collection of these data on a regular basis. Specific issues that can be addressed include changing patterns of union formation and dissolution, the situation of single parent families, and home leaving patterns of young adults.

A third but more minor theme of the cycle is concerned with the division of household labour.

### **5. PROGRAM OUTPUTS**

The GSS results are disseminated in a variety of ways. For each survey there are one or more publications that present the results of data analysis with respect to particular social issues and the monitoring of conditions and trends. The results of Cycle 1 are reported in Statistics Canada (1987) and Stone (1988); the results of Cycle 2 are presented in Harvey *et al.* (1991) and Creese *et al.* (1991); and the Cycle 3 results are reported in Sacco and Johnson (1990) and Millar and Adams (1991). Publications containing the results of other cycles are in preparation. The general public are made aware of GSS results through the publication of reports in the media which are often based on articles published in *Canadian Social Trends*, a quarterly Statistics Canada publication that is targeted to a general audience.



A second product is a public use microdata file and associated documentation to enable university and other researchers to carry out their own analysis of the data. These data are also useful for teaching purposes. Microdata files from the first five survey cycles are now available.

In addition to the product outputs, the GSS program has developed a survey capacity. This is not simply a system for data collection and processing, but includes other major components. Content research and development, related data specification, analysis of survey and other relevant data, dissemination of informative results as well as the development and use, where applicable, of improved methods of collection, processing, analysis and dissemination are all components of the evolving survey capacity of the GSS group.

## 6. FUTURE PLANS

As the GSS program moves into the second round of surveys, attention has shifted from the problems of developing and fielding five new surveys to further building the survey program through partnerships with others. The first round of surveys has had a modest success with obtaining buy-ins of additional sample and/or focus content. Only Cycle 4 had neither focus content nor increased sample size. For the first time, the 1990 survey had provincial participation, with the Ontario government funding an increase in sample size.

A new initiative for the GSS program is an investigation of the potential for expanding the scope of the survey to include interviewing a sub-sample of respondents again in future cycles. In the short term, this could provide an enriched data set by linking content from different cycles. In the longer term, it could serve to provide longitudinal data by interviewing respondents on the same topic five years later. A feasibility study was conducted in 1990 and the possibility of interviewing a sample of respondents from a previous cycle is now offered to interested sponsors.

The GSS will also continue to undertake a range of more general research and development activities. Core content of the first set of cycles will be reviewed and input sought from users as to possible improvements for future cycles. While new and alternative survey designs and approaches will be considered, any potential changes will have to be balanced against the impact on data comparability that is required for the long term goal of monitoring change. In addition, the content from the first round of surveys will be reviewed from the point of view of consistency and integration across GSS survey cycles and between the GSS and the 1991 Census and other household surveys. On-going development of the GSS infrastructure will also continue. Consideration was given to changing to monthly data collection (and monthly data collection was implemented for Cycle 6) and will be given to supplemental collection methods (*e.g.* mail). Attempts are also being made to shift processing of the survey to a micro-computer environment to further improve timeliness. Finally, new procedures, such as computer-assisted telephone interviewing, will be considered as these become available as part of a larger Statistics Canada survey development program.

In summary, the GSS Program during the coming years will focus on building on the firm foundation that has been established during the first round of surveys. The primary objective will continue to be the measurement of social conditions and the gradual development of a time series to monitor trends. In addition, flexibility will be maintained in order to quickly respond to new and emerging social information needs.

## REFERENCES

- BURCH, T.K. (1985). *Family History Survey: Preliminary Findings*. Catalogue 99-955, Statistics Canada.
- CREESE, G., GUPPY, N., and MEISSNER, M. (1991). Ups and downs on the ladder of success. *General Social Survey Analysis Series*. Catalogue 11-612E, No. 5, Statistics Canada.
- HARVEY, A.S., MARSHALL, K., and FREDERICK, J.A. (1991). Where does time go? *General Social Survey Analysis Series*. Catalogue 11-612E, No. 4, Statistics Canada.
- KISH, L. (1949). A procedure for objective respondent selection within the household. *Journal of the American Statistical Association*, 44, 380-387.
- KISH, L. (1976). Optima and proxima in linear sample designs. *Journal of the Royal Statistical Society A*, 139, 80-95.
- MILLAR, W., and ADAMS, O. (1991). Accidents in Canada. *General Social Survey Analysis Series*. Catalogue 11-612E, No. 3, Statistics Canada.
- SACCO, V., and JOHNSON, H. (1990). Patterns of criminal victimization in Canada. *General Social Survey Analysis Series*. Catalogue 11-612, No. 2, Statistics Canada.
- STATISTICS CANADA (1987). Health and social support, 1985. *General Social Survey Analysis Series*. Catalogue 11-612, No. 1, Statistics, Canada.
- STATISTICS CANADA (1989). *Household facilities and equipment*. Catalogue 64-202, Statistics Canada.
- STATISTICS CANADA (1990a). *Overview of Special Surveys 1989*, Household Surveys Division, Statistics Canada.
- STATISTICS CANADA (1990b). *Household facilities by income and other characteristics*. Catalogue 13-218, Statistics Canada.
- STONE, L. (1988). *Family and Friendship Ties Among Canada's Seniors*. Catalogue 89-508, Statistics Canada.
- WAKSBERG, J. (1978). Sampling methods for random digit dialing. *Journal of the American Statistical Association*, 73, 40-46.

## ACKNOWLEDGEMENTS

*Survey Methodology* wishes to thank the following persons who have served as referees, sometimes more than once, during 1991:

- |   |  |
|---|--|
| J. Alho, <i>University of Joensuu</i>                                     | B. Lefrançois, <i>Statistics Canada</i>                            |
| J. Armstrong, <i>Statistics Canada</i>                                    | T.P. Liu, <i>Statistics Canada</i>                                 |
| Y. Bélanger, <i>Statistics Canada</i>                                     | M. March, <i>Statistics Canada</i>                                 |
| D. Bellhouse, <i>University of Western Ontario</i>                        | S.M. Miller, <i>U.S. Bureau of Labor Statistics</i>                |
| K. Bennett, <i>Statistics Canada</i>                                      | I. Munck, <i>Statistics Sweden</i>                                 |
| J.-M. Berthelot, <i>Statistics Canada</i>                                 | J.C. Nash, <i>University of Ottawa</i>                             |
| D.A. Binder, <i>Statistics Canada</i>                                     | H.B. Newcombe, <i>Consultant</i>                                   |
| K. Bollen, <i>University of North Carolina – Chapel Hill</i>              | S. Presser, <i>University of Maryland</i>                          |
| P.A. Buesher, <i>North Carolina Center for Health Statistics</i>          | D. B. Radner, <i>U.S. Social Security Administration</i>           |
| K.P. Burnham, <i>U.S. Fish and Wildlife and Colorado State University</i> | J.N.K. Rao, <i>Carleton University</i>                             |
| S.J. Butani, <i>U.S. Bureau of Labor Statistics</i>                       | P.S.R.S. Rao, <i>University of Rochester</i>                       |
| G.H. Choudhry, <i>Statistics Canada</i>                                   | L.-P. Rivest, <i>Université Laval</i>                              |
| M.L. Cohen, <i>University of Maryland</i>                                 | W. Rodgers, <i>University of Michigan</i>                          |
| C.D. Cowan, <i>Opinion Research Corporation</i>                           | D.B. Rubin, <i>Harvard University</i>                              |
| E.B. Dagum, <i>Statistics Canada</i>                                      | K. Rust, <i>Westat Inc.</i>  |
| J.-C. Deville, <i>INSEE</i>   | I. Sande, <i>Bell Communications Research</i>                      |
| D. Dolson, <i>Statistics Canada</i>                                       | C.E. Särndal, <i>University of Montreal</i>                        |
| J.D. Drew, <i>Statistics Canada</i>                                       | A. Satin, <i>Statistics Canada</i>                                 |
| F.J. Fowler, Jr., <i>University of Massachusetts</i>                      | W.L. Schaible, <i>U.S. Bureau of Labor Statistics</i>              |
| J.G. Gambino, <i>Statistics Canada</i>                                    | F.J. Scheuren, <i>U.S. Internal Revenue Service</i>                |
| J.F. Gentleman, <i>Statistics Canada</i>                                  | I. Schiopu-Kratina, <i>Statistics Canada</i>                       |
| M.E. Gonzalez, <i>U.S. Office of Management and Budget</i>                | K.P. Srinath, <i>Statistics Canada</i>                             |
| J.-F. Gosselin, <i>Statistics Canada</i>                                  | C.M. Suchindran, <i>University of North Carolina – Chapel Hill</i> |
| H. Gough, <i>Statistics Canada</i>  | S. Sudman, <i>University of Illinois – Urbana-Champaign</i>        |
| A. Gower, <i>Statistics Canada</i>  | A. Sunter, <i>A.B. Sunter Research Design &amp; Analysis, Inc.</i> |
| G.B. Gray, <i>Statistics Canada</i>                                       | L. Swain, <i>Statistics Canada</i>                                 |
| M.A. Hidioglou, <i>Statistics Canada</i>                                  | R.B.P. Verma, <i>Statistics Canada</i>                             |
| M.A. Hill, <i>Systat Inc.</i>   | P.R. Voss, <i>University of Wisconsin – Madison</i>                |
| G.J.C. Hole, <i>Statistics Canada</i>                                     | J. Waksberg, <i>Westat Inc.</i>                                    |
| D. Holt, <i>University of Southampton</i>                                 | G.S. Werking, <i>U.S. Bureau of Labor Statistics</i>               |
| J. Hox, <i>University of Amsterdam</i>                                    | W.E. Winkler, <i>U.S. Bureau of the Census</i>                     |
| B. Hulliger-Domingues, <i>Swiss Federal Statistical Office</i>            | K.M. Wolter, <i>A.C. Nielsen</i>                                   |
| G. Kalton, <i>University of Michigan</i>                                  | A. Zaslavsky, <i>Harvard University</i>                            |

Acknowledgements are also due to those who assisted during the production of the 1991 issues: S. Beauchamp and S. Lineger (Photocomposition), G. Gaulin (Author Services), and M. Haight (Translation Services). Finally we wish to acknowledge M. Kent, C. Larabie and D. Lemire of Social Survey Methods Division, for their support with coordination, typing and copy editing.





# GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue (Vol. 10, No. 2 and onward) of Survey Methodology as a guide and note particularly the following points:

## 1. Layout

- 1.1 Manuscripts should be typed on white bond paper of standard size ( $8\frac{1}{2} \times 11$  inch), one side only, entirely double spaced with margins of at least  $1\frac{1}{2}$  inches on all sides.
- 1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.
- 1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.
- 1.4 Acknowledgements should appear at the end of the text.
- 1.5 Any appendix should be placed after the acknowledgements but before the list of references.

## 2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

## 3. Style

- 3.1 Avoid footnotes, abbreviations, and acronyms.
- 3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as “exp(·)” and “log(·)”, etc.
- 3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
- 3.4 Write fractions in the text using a solidus.
- 3.5 Distinguish between ambiguous characters, (e.g., w,  $\omega$ ; o, O, 0; 1, 1).
- 3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

## 4. Figures and Tables

- 4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.
- 4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

## 5. References

- 5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, p. 164).
- 5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

# DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Avant de dactylographier votre texte pour le soumettre, prière d'examiner un numéro récent de Techniques d'enquête (à partir du vol. 10, n° 2) et de noter les points suivants:

1.

Présentation

1.1

Les textes doivent être dactylographiés sur un papier blanc de format standard (8½ par 11 pouces), sur une face seulement, à double interligne partout et avec des marges d'au moins 1½ pouce tout autour.

1.2

Les textes doivent être divisés en sections numérotées portant des titres appropriés. Le nom et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.

1.4

Les remerciements doivent paraître à la fin du texte.

1.5

Toute annexe doit suivre les remerciements mais précéder la bibliographie.
2.

Résumé

Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.

3.

Rédaction

3.1

Éviter les notes au bas des pages, les abréviations et les sigles.

3.2

Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme exp( ) et log( ) etc.

3.3

Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin.

3.4

Écrire les fractions dans le texte à l'aide d'une barre oblique.

3.5

Distinguer clairement les caractères ambigus (comme w, ω; o, O; 1, l).

3.6

Les caractères italiques sont utilisés pour faire ressortir des mots. Indiquer ce qui doit être imprimé en italique en le soulignant dans le texte.

4.

Figures et tableaux

4.1

Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).

4.2

Ils doivent paraître sur des pages séparées et porter une indication de l'endroit où ils doivent figurer dans le texte. (Normalement, ils doivent être insérés près du passage qui y fait référence pour la première fois.)

5.

Bibliographie

5.1

Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence.

5.2

La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.





REMERCIEMENTS

Techniques d'enquête désire remercier les personnes suivantes, qui ont accepté de faire la critique d'un article, souvent plus d'une fois, durant l'année 1991:

J. Alho, *University of Joensuu*  
J. Armstrong, *Statistique Canada*  
Y. Bélanger, *Statistique Canada*  
D. Bellhouse, *University of Western Ontario*  
K. Bennett, *Statistique Canada*  
J.-M. Berthelot, *Statistique Canada*  
D.A. Binder, *Statistique Canada*  
K. Bollen, *University of North Carolina – Chapel Hill*  
P.A. Buesher, *North Carolina Center for Health Statistics*  
K.P. Burnham, *U.S. Fish and Wildlife et Colorado State University*  
S.J. Butani, *U.S. Bureau of Labor Statistics*  
G.H. Choudhry, *Statistique Canada*  
M.L. Cohen, *University of Maryland*  
C.D. Cowan, *Opinion Research Corporation*  
E.B. Dagum, *Statistique Canada*  
J.-C. Deville, *INSEE*  
D. Dolson, *Statistique Canada*  
J.D. Drew, *Statistique Canada*  
F.J. Fowler, Jr., *University of Massachusetts*  
J.G. Gambino, *Statistique Canada*  
J.F. Gentileman, *Statistique Canada*  
M.E. Gonzalez, *U.S. Office of Management and Budget*  
J.-F. Gosselin, *Statistique Canada*  
H. Gough, *Statistique Canada*  
A. Gower, *Statistique Canada*  
G.B. Gray, *Statistique Canada*  
M.A. Hidiroglou, *Statistique Canada*  
M.A. Hill, *Systat Inc.*  
G.J.C. Hole, *Statistique Canada*  
D. Holt, *University of Southampton*  
J. Hox, *University of Amsterdam*  
B. Hülliger-Domingues, *Office fédéral de la statistique, Suisse*  
G. Kalton, *University of Michigan*

B. Lefrançois, *Statistique Canada*  
T.P. Liu, *Statistique Canada*  
M. March, *Statistique Canada*  
S.M. Miller, *U.S. Bureau of Labor Statistics*  
I. Munck, *Statistics Sweden*  
J.C. Nash, *Université d'Ottawa*  
H.B. Newcombe, *Expert conseil*  
S. Presser, *University of Maryland*  
D. B. Radner, *U.S. Social Security Administration*  
J.N.K. Rao, *Carleton University*  
P.S.R.S. Rao, *University of Rochester*  
L.-P. Rivest, *Université Laval*  
W. Rodgers, *University of Michigan*  
D.B. Rubin, *Harvard University*  
K. Rust, *Westat Inc.*  
I. Sande, *Bell Communications Research*  
C.E. Särndal, *Université de Montréal*  
A. Satin, *Statistique Canada*  
W.L. Schaible, *U.S. Bureau of Labor Statistics*  
F.J. Scheuren, *U.S. Internal Revenue Service*  
I. Schiopu-Kratina, *Statistique Canada*  
K.P. Srinath, *Statistique Canada*  
C.M. Suchindran, *University of North Carolina – Chapel Hill*  
S. Sudman, *University of Illinois – Urbana-Champaign*  
A. Sunter, *A.B. Sunter Research Design & Analysis, Inc.*  
L. Swain, *Statistique Canada*  
R.B.P. Verma, *Statistique Canada*  
P.R. Voss, *University of Wisconsin – Madison*  
J. Wakeberg, *Westat Inc.*  
G.S. Wering, *U.S. Bureau of Labor Statistics*  
W.E. Winkler, *U.S. Bureau of the Census*  
K.M. Wolter, *A.C. Nielsen*  
A. Zaslavsky, *Harvard University*

On remercie également ceux qui ont contribué à la production des numéros de la revue pour 1991: S. Beauchamp et S. Liniger (Photocomposition), G. Gaulin (Services aux auteurs) et M. Haighi (Services de traduction). Finalement on désire exprimer notre reconnaissance à M. Kent, C. Larabie et D. Lemire de la Division des méthodes d'enquêtes sociales, pour leur apport à la coordination, la dactylographie et la rédaction.

- STATISTICS CANADA (1987). Santé et aide du milieu, 1985. *Enquête sociale générale Série analytique*. Catalogue n° 11-612F, n° 1, Statistique Canada.
- STATISTICS CANADA (1989). *L'équipement ménager*. Catalogue n° 64-202, Statistique Canada.
- STATISTICS CANADA (1990a). *Aperçu des enquêtes spéciales 1989*. Division des enquêtes-ménages, Statistique Canada.
- STATISTICS CANADA (1990b). *Équipement ménager selon le revenu et d'autres caractéristiques*. Catalogue n° 13-218, Statistique Canada.
- STONE, L. (1988). *Family and Friendship Ties Among Canada's Seniors*. Catalogue n° 89-508, Statistique Canada.
- WAKSBERG, J. (1978). Sampling methods for random digit dialing. *Journal of the American Statistical Association*, 73, 40-46.



L'équipe de l'ESG a également étudié la possibilité d'élargir la portée de l'enquête en interviewant de nouveau un sous-échantillon de répondants à l'occasion des cycles d'enquête ultérieurs. À court terme, cette mesure pourrait permettre d'obtenir un ensemble de données enrichi en couplant les données recueillies dans le cadre des différents cycles. À long terme, elle pourrait permettre d'établir une base de données longitudinales en interviewant les répondants sur le même sujet cinq ans plus tard. Une étude de faisabilité a été réalisée en 1990 et les parrains intéressés ont maintenant la possibilité de financer la réalisation d'interviews auprès d'un échantillon de répondants ayant participé à un cycle d'enquête antérieur.

Par ailleurs, l'équipe de l'ESG continuera de réaliser divers travaux de recherche et de développer de nature plus générale. On procédera à un examen de la composante de la thématique principale des cinq premiers cycles et les utilisateurs seront invités à proposer des améliorations susceptibles d'être apportées au cours des cycles ultérieurs. Même si on étudie la possibilité d'utiliser de nouveaux plans et méthodes d'échantillonnage, il faudra, avant de faire quelque modification que ce soit, tenir compte de l'effet de cette modification sur la comparabilité des données, caractéristique essentielle pour qu'il soit possible, conformément à l'objectif à long terme de l'enquête, de suivre l'évolution des tendances. En outre, l'équipe de l'ESG examinera le contenu des cinq premiers cycles de l'enquête afin d'en vérifier la cohérence et de déterminer le degré d'intégration entre les divers cycles de l'ESG ainsi qu'entre l'ESG, le recensement de 1991 et les autres enquêtes-ménages. Les travaux courants de développement de l'infrastructure de l'ESG se poursuivront aussi. Après avoir examiné la question, l'équipe a décidé de procéder à une collecte mensuelle des données à l'occasion du sixième cycle de l'enquête et elle étudiera la possibilité d'utiliser des méthodes de collecte complémentaires (*p. ex.*, retour par la poste). On tente actuellement de remplacer les mini-ordinateurs actuellement utilisés pour le traitement des données par des micro-ordinateurs afin de réduire davantage les délais d'exécution. Enfin, l'équipe étudiera la possibilité d'utiliser de nouvelles techniques, comme l'interview téléphonique assistée par ordinateur, à mesure qu'elles seront mises au point dans le cadre du programme d'élaboration des enquêtes de Statistique Canada.

Bref, au cours des prochaines années, l'équipe de l'ESG se propose de faire fond sur les assises solides établies au cours de la première série de cinq cycles d'enquêtes. L'enquête continuera d'avoir pour principal objectif la mesure des conditions sociales et l'élaboration graduelle d'une série chronologique permettant de suivre l'évolution des tendances. En outre, on s'assurera de maintenir la souplesse de l'enquête afin d'être en mesure de satisfaire rapidement aux nouveaux besoins en matière de statistique sociale.

## BIBLIOGRAPHIE

- BURCH, T.K. (1985). *Enquête sur la famille: conclusions préliminaires*. Catalogue n° 99-955, Statistique Canada.
- CREESE, G., GUPPY, N., et MEISSNER, M. (1991). Mobilité sociale ascendante et descendante au Canada. *Enquête sociale générale Série analytique*. Catalogue n° 11-612F, n° 5, Statistique Canada.
- HARVEY, A.S., MARSHALL, K., et FREDERICK, J.A. (1991). L'emploi du temps. *Enquête sociale générale Série analytique*. Catalogue n° 11-612F, n° 4, Statistique Canada.
- KISH, L. (1949). A procedure for objective respondent selection within the household. *Journal of the American Statistical Association*, 44, 380-387.
- KISH, L. (1976). Optima and proxima in linear sample designs. *Journal of the Royal Statistical Society A*, 139, 80-95.
- MILLAR, W., et ADAMS, O. (1991). Accidents au Canada. *Enquête sociale générale Série analytique*. Catalogue n° 11-612F, n° 3, Statistique Canada.
- SACCO, V., et JOHNSON, H. (1990). Profil de la victimisation au Canada. *Enquête sociale générale Série analytique*. Catalogue n° 11-612F, n° 2, Statistique Canada.

et familiaux. On a déjà procédé à la collecte rétrospective de telles données au moyen d'une enquête spéciale sur les antécédents familiaux réalisée en 1984 (Burch 1985). Le cycle de l'ESG portant sur la famille prévoit la collecte de telles données à intervalles réguliers. Figurent au nombre des questions pouvant être étudiées dans le cadre de ce cycle: l'évolution des tendances en matière de formation et de dissolution des unions, la situation des familles monoparentales et l'âge auquel les jeunes adultes ont tendance à quitter le foyer paternel.

Le troisième thème de ce cycle, qui occupe une place secondaire par rapport aux deux premiers, est celui de la répartition des travaux domestiques.

## 5. PRODUITS DU PROGRAMME

Les résultats de l'ESG sont diffusés de diverses façons. Chaque cycle d'enquête donne lieu à l'élaboration d'au moins une publication qui présente les résultats de l'analyse des données recueillies en regard à des questions sociales précises et qui brosse un tableau de la conjoncture et des tendances. On trouve un compte rendu des résultats du premier cycle de l'enquête dans Statistique Canada (1987) et dans Stone (1988); les résultats du deuxième cycle apparaissent dans Harvey et coll. (1991) ainsi que dans Creese et coll. (1991); enfin ceux du troisième cycle sont présentés dans Sacco et Johnson (1990) et dans Millar et Adams (1991). Les publications faisant état des résultats des autres cycles de l'enquête sont en cours d'élaboration. Le grand public est informé des résultats de l'ESG par la diffusion dans les médias de rapports qui sont souvent basés sur des articles parus dans *Tendances sociales canadiennes*, publication trimestrielle de Statistique Canada s'adressant au grand public.

L'équipe de l'ESG élabore également des fichiers de microdonnées et une documentation connexe afin de permettre aux chercheurs universitaires ou autres d'effectuer leur propre analyse des données. Les données comprises dans ces fichiers peuvent également être utilisées à des fins didactiques. On peut actuellement se procurer les fichiers de microdonnées relatifs aux quatre premiers cycles de l'enquête.

Outre ces produits, l'équipe de l'ESG a aussi élaboré un système d'enquête qui ne se résume pas à un système de collecte et de traitement des données, mais comprend d'autres composantes majeures. Figurent au nombre de ces composantes: la recherche et le développement en matière de contenu, les spécifications relatives aux données connexes, l'analyse des données d'enquête et d'autres données pertinentes, la diffusion des résultats ainsi que l'élaboration et l'utilisation, lorsqu'il y a lieu, de méthodes améliorées de collecte, de traitement, d'analyse et de diffusion des données.

## 6. PERSPECTIVES D'AVENIR

Avec le début de la deuxième série de cycles d'enquête, l'équipe de l'ESG a détourné son attention des problèmes soulevés par l'élaboration et la mise en oeuvre de cinq nouvelles enquêtes pour axer davantage ses efforts sur le développement ultérieur du programme par le biais d'accords de partenariat. Au cours de la première série de cycles d'enquête, les efforts déployés pour amener des tiers à financer un élargissement de la taille de l'échantillon et/ou la mise en oeuvre d'une composante de la thématique particulière ont été couronnés d'un certain succès. Seul le quatrième cycle de l'enquête n'a pas comporté de composante de la thématique particulière et n'a donné lieu à aucun élargissement de la taille de l'échantillon. Le cycle de 1990 a pour sa part été marqué par la première participation d'une administration provinciale à l'enquête, le gouvernement de l'Ontario ayant alors financé une augmentation de la taille de l'échantillon.



conséquences sur le plan des limitations d'activité, sur les soins médicaux reçus et sur les pertes financières subies. En outre, on demandait aux répondants d'indiquer leur perception eu égard au risque d'être victime d'un accident ou d'un acte criminel ainsi que les précautions prises pour éviter l'occurrence de tels incidents.

La composante de la thématique particulière de ce cycle comportait une série de questions sur le contact avec le système de justice pénale (p. ex., police, tribunaux, avocats) ainsi que sur la connaissance et l'utilisation des services offerts aux victimes d'actes criminels.

#### 4.4 Études et travail

Bien que l'Enquête mensuelle sur la population active et les autres enquêtes relatives au travail fournissent un trésor de renseignements sur la population active, aucune des enquêtes existantes ne recueille beaucoup de données sur les aspects sociaux du travail ou sur la qualité de la vie active. Le cycle de l'ESG portant sur les études et le travail, qui a été réalisé au début de 1989, a été spécialement conçu pour combler partiellement cette lacune.

Ce cycle s'articule autour de trois grands thèmes qui reflètent certains des changements fondamentaux que connaît la société canadienne: les tendances sur le plan du travail et des études; les nouvelles technologies et les ressources humaines; enfin, le travail dans le secteur des services. Ces thèmes sont représentatifs d'un large éventail de questions sur lesquelles il faut recueillir plus de renseignements. Ainsi, le rythme de plus en plus rapide de l'innovation technologique exige l'acquisition d'une connaissance détaillée de l'utilisation des ordinateurs ainsi que l'obtention d'une formation en la matière; l'utilisation efficace du capital humain de la nation nécessite une meilleure compréhension des liens qui existent entre la population active et le système d'éducation; nous nous devons également de prévoir les besoins auxquels devront satisfaire les établissements d'enseignement ainsi que l'évolution des liens entre le niveau de scolarité et le statut socio-économique. Ce cycle de l'ESG vient aussi compléter les sources de données existantes en recueillant de nouvelles données sur la population âgée ainsi que sur certaines implications socio-économiques de l'atteinte de l'âge moyen par la génération du baby-boom.

Ce cycle d'enquête a permis de recueillir des données partielles sur les antécédents professionnels et scolaires ainsi que des données sur la formation technique, l'utilisation des ordinateurs et les perspectives d'avenir en matière de scolarité. Il a également permis de recueillir des données subjectives grâce à une série de questions sur le degré de satisfaction éprouvé eu égard à la retraite et à d'autres aspects de l'existence, ainsi qu'à un bloc de questions sur les attitudes face à la science et à la technologie.

#### 4.5 Famille et amis

Le cinquième cycle de l'ESG, réalisé au début de 1990, portait sur diverses questions relatives à la famille et aux amis. Bien que le recensement et les autres enquêtes-ménages fournissent des données sur la famille, l'évolution de la vie familiale a nécessité la collecte de nouveaux types de renseignements. Une des lacunes de données existantes tient au fait qu'elles sont généralement basées sur un concept plutôt étroit de la famille, considérée comme une famille nucléaire regroupant parents et enfants ou comme une famille économique constituée de personnes apparentées faisant partie du même ménage. Ce cycle d'enquête étudie la famille dans un contexte plus large et permet de recueillir des données sur l'étendue et la nature des réseaux d'aide fournie par des proches ainsi que sur les questions connexes des formes que prennent l'aide et le soutien spontanés fournis par la famille et les amis.

Le deuxième grand thème de ce cycle d'enquête a été choisi en raison des tendances observées sur le plan du mariage et du divorce ainsi que de la fréquence accrue des unions de fait. Un nombre de plus en plus élevé de Canadiens connaissent plus d'une union au cours de leur vie. Ces changements ont, sur la vie familiale et sur les enfants, une incidence considérable qui ne saurait être mieux cernée qu'à l'aide d'une analyse des données sur les antécédents matrimoniaux



La composante de la thématique particulière du premier cycle de l'ESG était axée sur l'aide apportée aux personnes âgées et portait sur les réseaux sociaux, le soutien donné et reçu, ainsi que la participation à diverses activités sociales. On a accru la taille de l'échantillon de personnes âgées afin de permettre l'analyse plus approfondies.

#### 4.2 Emploi du temps

La composante de la thématique principale du deuxième cycle de l'ESG était l'emploi du temps, tel que déterminé à partir du "journal d'une journée", généralement celle précédant le jour de l'interview. Les répondants indiquaient leurs activités principales au cours de cette journée, l'heure du début et de la fin de ces activités, l'endroit où elles avaient eu lieu et les personnes en compagnie desquelles elles avaient été exercées (p. ex., époux(se), enfants ou amis). Ces données ont permis d'obtenir des renseignements sur la fréquence avec laquelle les gens participent à des activités comme le travail rémunéré, les travaux domestiques, l'assistance à des activités culturelles et le visionnement de la télévision, ainsi que sur le temps qu'ils consacrent à ces activités.

L'enquête indique de quelle façon les Canadiens répartissent leur temps entre des activités comme le travail rémunéré, les travaux domestiques et les autres genres de travail non rémunéré, ainsi que les activités de loisir. On peut utiliser les données pour souligner les contraintes auxquelles est assujéti l'emploi du temps d'une personne et indiquer comment ces contraintes sont réparties entre divers groupes de population. L'inclusion d'une série de questions sur le degré de satisfaction éprouvé à l'égard de divers aspects de l'existence permet en outre d'établir une corrélation entre les mesures ainsi obtenues et l'emploi du temps pour divers groupes de population.

Le cycle de l'ESG de 1986 sur l'emploi du temps comportait aussi un petit module sur la capillarité sociale, qui permet d'analyser le mouvement du répondant par rapport à ses parents sous l'angle de la mobilité scolaire ou professionnelle.

La composante de la thématique particulière de ce cycle comportait un ensemble détaillé de questions sur la connaissance et l'utilisation de la langue. Bien qu'on s'attende généralement à ce qu'il existe un lien entre la thématique particulière et la thématique principale et à ce que la première vienne compléter la seconde, il existait aussi une demande pour des données sur la langue beaucoup plus détaillées que celles qu'il est possible de recueillir dans le cadre du recensement de la population. Les données recueillies avaient trait, entre autres, à la langue utilisée à diverses époques de la vie (p. ex., première langue apprise, pendant l'enfance, à l'école) et dans divers contextes, comme à la maison, au travail, avec les amis, en visionnant la télévision et dans les communications avec les organismes fédéraux. Afin de permettre une analyse plus approfondie dans les régions bilingues du pays, on avait accru la taille de l'échantillon prélevé dans ces régions.

#### 4.3 Risques personnels

La composante de la thématique principale du troisième cycle de l'ESG avait pour sujet les risques personnels, qu'il s'agisse des risques d'être victime d'un accident ou d'un acte criminel. Traditionnellement, les renseignements sur ces sujets ont été obtenus de sources administratives, comme les statistiques policières et les dossiers d'hôpital. Or, non seulement ces sources ne fournissent-elles que très peu de renseignements sur les victimes, mais nombre de crimes (plus de la moitié selon les estimations de l'ESG) et d'accidents ne sont pas déclarés aux autorités.

Le cycle de l'enquête sur les risques personnels, réalisé au début de 1988, portait plus particulièrement sur les accidents et les actes criminels dont les répondants avaient été victimes l'année précédente (1987). Des données ont également été recueillies sur les habitudes de vie, comme la consommation d'alcool et la fréquence des sorties nocturnes, afin d'établir une corrélation entre ces habitudes et le fait d'être victime d'un accident ou d'un acte criminel. Pour chaque accident ou acte criminel déclaré, on recueillait des données sur la nature de l'incident, sur ses

où :

$n_R$  est le nombre de ménages répondants dans la "strate" de non-réponse, et  $n_{NR}$  est le nombre correspondant de ménages non répondants.

### 3.11.3 Ajustement en fonction du nombre de téléphones

Comme les ménages ayant plus d'un numéro de téléphone ont une plus forte probabilité d'être échantillonnés dans le cadre d'une enquête téléphonique par CA, le poids initial ajusté en fonction de la non-réponse (poids attribué aux numéros de téléphone) fait l'objet d'un nouvel ajustement consistant à le diviser par le nombre de numéros de téléphone que possède le ménage afin d'obtenir un poids du ménage.

#### 3.11.4 Poids initial des personnes

Comme on n'interview qu'un seul répondant admissible par ménage, il faut ajuster le poids du ménage en le multipliant par le nombre de répondants admissibles afin d'obtenir un poids de la personne.

#### 3.11.5 Stratification a posteriori

Enfin, on utilise les projections démographiques du recensement comme totaux de référence pour la stratification a posteriori des poids des personnes, premièrement en fonction de la taille estimative de la population de la strate, puis en fonction de la taille estimative de l'effectif des groupes province-âge-sexe. (On notera que c'est seulement après la première étape de la stratification a posteriori que la somme des poids attribués dans les strates où l'échantillon a été prélevé par la méthode de Wakhsberg correspond à une estimation de la taille de la population. Jusqu'à cette étape, ils s'écartent d'un ensemble de poids basés sur l'inverse de la probabilité de tirage par une constante de proportionnalité inconnue.) Ces deux ensembles de totaux de référence sont ensuite utilisés comme bornes pour un ajustement des poids par la méthode itérative du quotient.

## 4. THEMES ET SUJETS DE RECHERCHE ABORDES

Comme nous l'avons vu plus haut, afin de permettre l'étude d'un large éventail de questions sociales, l'ESG porte sur un sujet différent de la thématique principale chaque année pendant cinq ans, puis étudie les mêmes sujets à intervalles réguliers. Ces sujets, qui ont été choisis afin de combler les lacunes apparentes du programme de la statistique sociale, sont exposés de façon plus détaillée ci-après.

### 4.1 Santé

La composante de la thématique principale du cycle portant sur la santé vise à fournir une gamme de mesures de l'état de santé telles que l'incapacité de courte et de longue durée, les problèmes de santé chroniques répandus, comme l'hypertension artérielle et le diabète, ainsi que l'utilisation des divers services de santé. De plus, elle permet de recueillir des données sur divers sujets liés au mode de vie, comme l'usage du tabac, la consommation d'alcool et l'activité physique. Une fois couplées aux données sur l'état de santé, ces données permettent d'obtenir des renseignements sur les obstacles que doivent franchir (p. ex., usage du tabac et consommation d'alcool) et sur les habitudes que doivent acquérir (p. ex., activité physique) les membres de divers groupes de population pour améliorer leur santé.

entre les taux de réponse enregistrés pour les ménages où aucun appel n'a été pris par un répondant et les taux enregistrés pour les ménages où au moins un des appels a été pris par un répondant. Bien que ce tableau n'indique aucun effet important de l'utilisation des répondants, il révèle un accroissement spectaculaire de la proportion des contacts établis au moyen d'un tel appareil (de 6,4% à 9,4% des ménages). Pour sa part, le tableau 4 établit une comparaison entre les taux de réponse enregistrés pour les ménages où le premier appel ayant reçu une réponse a été pris par un répondant et les taux enregistrés pour les autres ménages. Selon ce tableau, il semble qu'on enregistre des taux de réponse plus élevés pour les ménages où le premier contact a été établi avec un répondant. En revanche, rien ne semble indiquer que l'utilisation des répondants entraîne une baisse importante des taux de réponse.

### 3.10 Saisie et traitement des données

Les données des cinq cycles d'enquête ont été saisies directement sur mini-ordinateur dans les bureaux régionaux de Statistique Canada. Bien que certaines vérifications simples visant à déterminer la validité des données aient été effectuées au moment de la saisie, dans la plupart des cas, il était possible d'annuler ces vérifications à l'aide de fonctions spéciales. Après avoir été transmises à Ottawa, les données brutes ont été soumises à un ensemble complet de vérifications visant à repérer, et à corriger dans la mesure du possible, les réponses invalides ou incohérentes. Lorsqu'une réponse était manquante, invalide ou incohérente par rapport aux autres réponses et qu'il était impossible de déterminer la valeur appropriée à partir des autres réponses figurant sur le questionnaire, on attribuait à cette réponse le code "inconnu". Les seules exceptions à cette règle concernaient les trois variables nécessaires pour les fins de la pondération, à savoir l'âge, le sexe et le numéro de téléphone. Lorsque ces variables étaient manquantes, on consultait le questionnaire pour faciliter l'imputation de valeurs appropriées.

### 3.11 Pondération

#### 3.11.1 Poids initiaux

La méthode de Wakseberg et la méthode EBI produisent toutes deux des échantillons auto-pondérés de numéros de téléphone résidentiels. Bien que la méthode de Wakseberg ne fournisse pas d'estimation de ce poids, il est suffisant, pour les fins de la pondération des données de l'ESG, d'attribuer un poids initial de un (1) aux numéros de téléphone prélevés dans les strates où cette méthode a été utilisée. Dans les strates où l'échantillon a été prélevé à l'aide de la méthode EBI, le poids initial est égal à la réciproque de la probabilité de tirage du numéro de téléphone. Cette probabilité se définit simplement par :

$$\frac{n_c}{100 \times N_B},$$

où :

$n_c$  est le nombre de numéros de téléphone prélevés, et  
 $N_B$  le nombre de banques actives dans la base de sondage.

#### 3.11.2 Ajustement en fonction de la non-réponse

On ajuste le poids initial en fonction de la non-réponse en utilisant des "strates" d'ajustement définies à partir de la géographie téléphonique. De façon typique, il s'agit de banques pour l'échantillon obtenu par la méthode de Wakseberg et d'ensembles IRP, pour l'échantillon obtenu par la méthode EBI. Les poids initiaux sont augmentés par le facteur suivant :

$$\frac{n_R}{n_R + n_{NR}},$$



Tableau 2

Taux de réponse par cycle et par bureau régional  
(Les résultats pour le cycle 6 sont des résultats préliminaires -  
Le symbole désigne les bureaux ne réalisant pas d'interviews)

Bureau régional	Cycle					
	3	4	5	6		
St. John's	84.1	82.8	90.9	-		
Halifax	84.7	84.1	85.9	82		
Montréal	83.0	79.6	81.2	82		
Sturgeon Falls	76.5	81.1	71.5	71		
Toronto	87.0	75.4	63.0	-		
Winnipeg	84.3	87.0	84.3	89		
Edmonton	83.2	79.4	76.8	-		
Vancouver	75.3	80.2	79.6	82		
Canada	82.9	80.7	75.8	81		
Canada (sans Toronto)	82.1	81.8	80.1	-		

Tableau 3

Taux de réponse et de non-réponse (%) selon le genre de non-réponse et selon qu'un des appels ait été ou non pris par un répondant

Est-ce que certains appels ont été pris par un répondant?					
	Cycle 4		Cycle 5		
	Non	Oui	Non	Oui	
Refus du ménage	7.18	8.17	10.34	9.74	
Autres cas de non-réponse du ménage	6.45	5.89	7.19	7.41	
Refus de l'enquête	5.05	4.55	4.39	3.27	
Autres cas de non-réponse de l'enquête	1.68	2.01	2.31	3.15	
Réponses	79.64	79.38	75.76	76.44	
Nombre d'enregistrements	10,981	747	16,611	1,715	
	(93.6%)	(6.4%)	(90.6%)	(9.4%)	

Tableau 4

Taux de réponse et de non-réponse (%) selon le genre de non-réponse et selon le genre de contact initial

Est-ce que le premier contact a été établi avec un répondant?					
	Cycle 4		Cycle 5		
	Non	Oui	Non	Oui	
Refus du ménage	7.24	7.19	10.46	7.90	
Autres cas de non-réponse du ménage	6.46	5.40	7.15	8.06	
Refus de l'enquête	5.08	3.96	4.38	3.07	
Autres cas de non-réponse de l'enquête	1.70	1.80	2.43	1.92	
Réponse	79.52	81.65	75.58	79.05	
Nombre d'enregistrements	11,172	556	17,023	1,303	
	(95.3%)	(4.7%)	(92.9%)	(7.1%)	

Tableau 1  
Taux de réponse et de non-réponse (%) par cycle, selon le genre

Résultats	Cycle				
	1	2	3	4	5
Refus du ménage	6.2	6.2	6.0	7.2	10.3
Autres cas de non-réponse du ménage	4.4	6.8	6.6	6.4	7.2
Refus de l'enquête	1.3	2.8	1.3	1.7	2.4
Autres cas de non-réponse de l'enquête	4.8	3.5	3.2	3.9	4.3
Cas spéciaux de non-réponse au cycle 2		1.9			
Réponse	83.4	78.9	82.9	80.7	75.8

Le tableau 1 indique les taux de réponse enregistrés pour les cinq premiers cycles de l'enquête. Les catégories "Autres cas de non-réponse du ménage" et "Autres cas de non-réponse de l'enquête" englobent les cas où il a été impossible de réaliser l'interview en raison de problèmes linguistiques, d'une maladie, du décès d'un des membres de la famille et de l'absence du ménage ou du répondant pendant la période d'enquête; il ne fait aucun doute que certains de ces cas de non-réponse constituent des refus déguisés. Dans tous les cycles sauf le deuxième, on a réalisé les interviews le plus tôt possible après avoir joint les ménages. À l'occasion du cycle 2, un intervalle d'un mois s'est écoulé entre le contact initial avec les ménages et les interviews; une partie de la non-réponse peut être attribuée directement à cet intervalle. Selon le tableau 1, les taux de réponse semblent avoir tendance à diminuer du premier au cinquième cycle.

Il suffit cependant de jeter un coup d'oeil au tableau 2, qui indique les taux de réponse enregistrés dans les divers bureaux régionaux pour les cycles 3 à 6, pour constater que la situation n'est pas aussi simple, nombre de bureaux (Halifax, Montréal, Winnipeg) affichant des taux de réponse relativement stables pour ces cycles d'enquête. De fait, si on exclut les résultats obtenus pour Toronto, il s'avère que le taux de réponse n'a accusé qu'un léger recul entre les cycles 3 et 5. Nous avons remarqué que les intervieweurs d'expérience avaient tendance à obtenir des taux de réponse plus élevés. Il se peut que la variation spectaculaire des taux de réponse enregistrée sur l'ensemble de ces trois cycles par le bureau de Toronto soit attribuable dans une large mesure à la difficulté qu'il y a à embaucher et à retenir le personnel dans une ville dont l'économie est en plein essor. Il est également possible qu'une partie de cette variation soit attribuable à un changement survenu au sein de la population échantillonnée par le bureau de Toronto.

Les résultats préliminaires relatifs aux huit premiers mois (janvier à août 1991) de collecte du cycle 6 (voir le tableau 2) indiquent qu'il a été possible d'inverser cette tendance vers une diminution des taux de réponse. On a fait entre les cycles 5 et 6 un certain nombre de changements, dont les plus importants sont le passage à une collecte mensuelle des données et la réaffectation de l'échantillon relatifs aux bureaux ne participant pas à la collecte de données du cycle 6, les échantillons de St. John's, Toronto et Edmonton étant respectivement transférés à Halifax, Sturgeon Falls et Winnipeg.

Au cours de la collecte des données du cycle 3, les intervieweurs ont constaté un accroissement du nombre d'appels pris par des répondeurs téléphoniques. Il était donc possible que les répondants utilisent ces appareils pour filtrer leurs appels et que cette pratique provoque une augmentation des taux de non-réponse. Nous ne sommes pas capable de déterminer quels appels ont été pris par un répondeur dans le cadre des cycles 1 à 3, mais il nous est possible de le faire pour les cycles subséquents, ce qui nous permet d'analyser dans une certaine mesure l'effet de l'utilisation de ces appareils sur les taux de réponse. On trouve au tableau 3 une comparaison

Lorsqu'on dispose uniquement d'une liste d'ensembles IRP, l'échantillon de l'ESP est prélevé au moyen de la méthode de Waksberg. Selon cette méthode, la probabilité de sélection des banques est proportionnelle à leur taille, laquelle correspond au nombre de numéros de téléphone résidentiels qu'elles comprennent. On prélève ensuite au sein de chacune des banques sélectionnées un échantillon aléatoire simple de numéros résidentiels. Lorsque la taille de l'échantillon est la même dans chaque banque, cette méthode permet d'obtenir un échantillon équiprobabiliste de numéros de téléphone résidentiels. Pour les fins de l'ESG, la taille de l'échantillon prélevé dans chaque banque est de 6 numéros. Cette méthode a pour avantage d'accroître la probabilité de tirage des numéros résidentiels, mais elle a pour inconvénient de produire un échantillon par grappes. Ainsi, dans certaines régions rurales de l'ouest canadien, seulement 6% environ des numéros générés à l'aide des listes IRP sont des numéros résidentiels, contre 50% des numéros prélevés au deuxième degré. Pour nombre de variables, les effets du plan attribuable à l'échantillonnage par grappes sont minimes, de l'ordre de 1.0 à 1.3.

Lorsqu'on dispose de renseignements plus détaillés permettant de dresser une liste de banques renfermant au moins un numéro résidentiel ("banques actives"), on utilise la méthode fondée sur l'élimination des banques inactives (EBI). Cette méthode consiste à prélever dans les banques actives un échantillon aléatoire simple de numéros de téléphone puis à rejeter les numéros non résidentiels afin d'obtenir un échantillon aléatoire simple de numéros résidentiels. Depuis le premier cycle de l'ESG en 1985, on a utilisé la méthode EBI pour une proportion de plus en plus grande de l'échantillon, à mesure qu'on pouvait obtenir plus de renseignements des compagnies de téléphone. À l'occasion du cycle 6 (mis en oeuvre en 1991), tout l'échantillon a été tiré à l'aide de cette méthode.

Un système de programmes informatiques a été rédigé à l'intention des bureaux régionaux de Statistique Canada afin de mettre en oeuvre ces deux méthodes d'échantillonnage et de suivre l'état d'avancement de l'enquête. Il faut que la totalité de l'échantillon relatif à une strate soit prélevé au moyen de la même méthode d'échantillonnage.

Une fois qu'un ménage a été joint par téléphone, on dresse une liste sur laquelle figurent le nom et l'âge de tous les membres du ménage, puis on choisit, au moyen de cette liste et d'un ensemble de numéros aléatoires imprimés pour chaque questionnaire, une personne âgée de 15 ans ou plus devant être interviewée. C'est la méthode de Kish (1949).

### 3.8 Échantillons spéciaux

Les parrains de l'ESG ont la possibilité de financer la réalisation d'interviews additionnelles. Les échantillons additionnels nécessaires peuvent être obtenus en élargissant la taille de l'échantillon CA prélevé dans une ou plusieurs strates ou encore être tirés dans d'autres bases de sondage.

À l'occasion des cycles 2 et 5, on a choisi d'élargir la taille des échantillons CA prélevés dans certaines strates présentant un intérêt particulier pour les parrains.

À l'occasion des cycles 1, 5 et 6, on a complété l'échantillon CA à l'aide d'échantillons additionnels de personnes appartenant à certains groupes d'intérêt. Dans chacun de ces cas, ces échantillons étaient composés de personnes âgées de 65 ans ou plus tirées à partir d'une liste des ménages ayant récemment fait partie de l'échantillon de l'EPA.

### 3.9 Taux de réponse

Un des inconvénients des enquêtes téléphoniques tient au fait que les répondants semblent trouver plus facile de refuser de participer à une telle enquête qu'à une enquête utilisant la technique de l'interview sur place. Les entreprises utilisent régulièrement le démarchage téléphonique pour vendre leurs produits et services, et tous ont appris à dire non au téléphone. De plus, les nouvelles technologies, comme les répondeurs téléphoniques et les fonctions spéciales dont sont pourvus les systèmes téléphoniques, offrent aux gens la possibilité de filtrer facilement leurs appels d'arrivée.



à l'Île-du-Prince-Édouard, deux à Terre-Neuve, en Nouvelle-Écosse, au Nouveau-Brunswick, au Manitoba, en Saskatchewan, en Alberta et en Colombie-Britannique, ainsi que trois au Québec et en Ontario.

Cette stratification de base peut cependant être modifiée, et l'a été à l'occasion, en fonction du caractère particulier du sujet traité ou des besoins spéciaux des parrains. À l'occasion du cycle 2, l'intérêt spécial porté à la langue utilisée a amené l'équipe de l'ESG à créer des strates distinctes et à utiliser un taux d'échantillonnage plus élevé dans les "régions de contact" où on croyait que résidaient un grand nombre d'anglophones et de francophones. De même, certaines strates spéciales ont été définies à l'occasion du cycle 5 afin de produire, à la demande d'un client, des estimations pour certaines régions intraprovinciales de l'Ontario.

3.6 Répartition de l'échantillon

La taille cible de l'échantillon de l'ESG est de 10,000 personnes (entreviws réalisées). Cet échantillon a été réparti entre les provinces en proportion de la racine carrée de leur population. Par après, l'échantillon provincial a été réparti entre les strates en proportion de la taille de ces dernières. La répartition proportionnelle à la racine carrée permet d'accroître la taille de l'échantillon pour les provinces plus petites (par rapport à l'échantillon qu'on obtiendrait au moyen d'une répartition proportionnelle à la population) sans compromettre la précision des estimations à l'échelle du Canada autant que le ferait une répartition directement proportionnelle. On a étudié la possibilité d'utiliser la méthode de Kish (1976) pour obtenir une répartition qui permette d'établir des estimations précises tant à l'échelle provinciale qu'à celle du pays, mais la répartition résultante n'a permis d'améliorer la précision des estimations à l'échelle nationale que dans une faible mesure tout en provoquant une modification spectaculaire, et jugée peu souhaitable, de la taille de l'échantillon de certaines provinces.

3.7 Méthode d'échantillonnage téléphonique

À l'exception des échantillons supplémentaires de personnes âgées de plus de 65 ans prélevés à partir des listes de ménages interviewés dans le cadre de l'Enquête sur la population active, les échantillons de l'ESG ont été prélevés au moyen des techniques de composition aléatoire. Deux techniques distinctes ont été employées, la méthode de Wakseberg (1978) et la méthode fondée sur l'élimination des banques inactives. Les deux méthodes utilisent les renseignements fournis par les compagnies de téléphone afin d'accroître la probabilité de joindre les ménages. C'est le niveau de détail des renseignements dont on dispose qui détermine la méthode utilisée.

Au Canada, les numéros de téléphone comportent dix chiffres que l'on peut décomposer en un "indicatif régional" de trois chiffres, un "préfixe" de trois chiffres et deux champs de deux chiffres, dont le premier porte le nom d'identificateur de banque. Ainsi, chaque ensemble défini par un "indicatif régional et un préfixe" (IRP) peut comporter dix mille numéros différents et chaque ensemble défini par un "indicatif régional, un préfixe et un identificateur de banque" (IRP – Banque ou simplement banque) peut comporter cent numéros différents. On trouve ci-après un numéro de téléphone fictif décomposé en ses divers composants:

216	Indicatif régional
357	Préfixe (central)
46	Identificateur de banque
75	Numéro
216-357	IRP
216-357-46	IRP – Banque (banque).

Lorsqu'on utilise la méthode de l'interview téléphonique, les personnes faisant partie de ménages qui n'ont pas le téléphone sont exclus de l'échantillon. Moins de 2% des ménages canadiens faisant partie de la population cible de l'Enquête sur la population active se trouvent dans cette situation (Statistique Canada 1989, 1990b). Toutefois, ce taux élevé de pénétration du téléphone n'est pas uniforme d'un groupe d'âge, d'une tranche de revenu, ni d'une province à l'autre: 95,4% des ménages de l'Île-du-Prince-Édouard ont le téléphone, contre 99,2% de ceux de l'Ontario; 99,1% des ménages dont le revenu s'établit entre 20 et 25 mille dollars ont le téléphone, contre seulement 93,9% de ceux dont le revenu est inférieur à 10 mille dollars. De plus, dans certaines sous-populations, le pourcentage de personnes possédant le téléphone est de loin inférieur à la moyenne; ainsi, ce pourcentage est seulement de 86,7% chez les personnes à faible revenu âgées de moins de 65 ans qui vivent seules.

Comme, de façon générale, les réponses par personne interposée ne sont pas acceptées pour les fins de l'ESG, les personnes incapables d'utiliser le téléphone (soit qu'elles soient sourdes ou muettes), ne pouvant être jointes au cours de la période d'enquête ou ne parlant ni anglais ni français sont exclues de la population étudiée. (Pour les fins du sixième cycle de l'ESG, qui porte sur la santé des Canadiens, on a décidé d'accepter les réponses par personnes interposées dans les cas où le répondant échantillonné ne peut faire l'interview en raison d'un problème de santé.)

Lorsque des échantillons supplémentaires sont tirés à partir de listes de ménages interviewés dans le cadre de l'Enquête sur la population active (comme dans le cas des 1<sup>er</sup>, 5<sup>e</sup> et 6<sup>e</sup> cycles de l'enquête), les habitants des réserves indiennes et les membres permanents des Forces armées canadiennes sont exclus de ces échantillons. Ces exclusions représentent moins de 0,5% des personnes âgées de plus de 65 ans (ce groupe d'âge est le seul au sein duquel de tels échantillons ont été prélevés).

### 3.5 Stratification

La stratification de l'échantillon de l'ESG tient compte des exigences relatives à l'estimation, des exigences opérationnelles, des restrictions que l'échantillonnage par composition aléatoire (CA) impose en matière de définition des strates, des problèmes de pondération soulevés par l'échantillonnage par composition aléatoire (CA) et des besoins spéciaux des parrains. Comme il est nécessaire d'établir certaines estimations à l'échelle provinciale, les limites des strates de l'ESG ne croisent jamais les limites provinciales. De même, les limites des strates ne croisent jamais les limites du secteur de compétence des bureaux régionaux car, pour des raisons opérationnelles, les répondants d'une même strate doivent être interviewés à partir d'un seul bureau régional. Par ailleurs, la méthode d'échantillonnage par composition aléatoire utilisée exige que les strates soient définies comme des groupes de centraux téléphoniques. Comme il faut, au moment de la pondération, disposer d'estimations précises de la taille des strates, les limites des strates (définies comme des groupes de centraux téléphoniques) doivent correspondre étroitement aux limites de groupes d'unités pour lesquels on dispose de données ou d'estimations précises sur la population. Pour les années intercensitaires, on dispose de telles données à l'échelle de la région métropolitaine de recensement (RMR).

La stratification de base établie en tenant compte de ces exigences a d'abord consisté à définir des strates dont les limites correspondaient aux limites des provinces. Étant donné qu'au cours des cycles 1 à 5 la Saskatchewan et l'Ontario relevaient chacune de deux bureaux régionaux, elles ont toutes deux été divisées en deux par une limite de strate. Dans un deuxième temps, les régions ainsi définies ont été chacune divisées en deux strates, l'une regroupant les RMR et l'autre, les autres régions. Enfin, les deux plus grandes RMR, celles de Montréal et de Toronto, ont été constituées en deux strates distinctes. Ainsi, l'échantillon des cycles 1 à 5 comportait 25 strates réparties comme suit: une à l'Île-du-Prince-Édouard (cette province ne comprend pas de RMR), deux à Terre-Neuve, en Nouvelle-Écosse, au Nouveau-Brunswick, au Manitoba, en Alberta et en Colombie-Britannique, trois au Québec, quatre en Saskatchewan et cinq en Ontario. L'échantillon du cycle 6 comportait 21 strates réparties comme suit: une



On peut conclure à la lumière de la dernière exigence que le plan d'échantillonnage se doit d'être simple puisque, en général, il est impossible de donner dans des fichiers à grande diffusion les renseignements sur le plan d'échantillonnage nécessaires pour analyser les données d'une enquête complexe. De même, la troisième exigence laisse entendre qu'il faut éviter de trop optimiser le plan d'échantillonnage pour des variables précises.

### 3.2 Mode de collecte des données

Il a fallu, au moment du choix du mode de collecte, tenir compte d'un certain nombre de facteurs concurrents: coût de l'interview, durée de l'interview, taux de réponse, exactitude des données recueillies et taille de l'échantillon. Compte tenu du niveau de détail des données devant être recueillies, on prévoyait une durée d'interview de 20 à 30 minutes par répondant. Afin de réduire le fardeau de déclaration du ménage et d'éviter tout effet de grappe au niveau de ce dernier, on a décidé de n'interviewer qu'une seule personne par ménage. Les principales méthodes de collecte qu'on a étudié la possibilité de retenir ont été: le retour par la poste, l'interview sur place et l'interview téléphonique. Compte tenu du caractère hétérogène de la population cible, on a jugé que les taux élevés de non-réponse obtenus avec la méthode du retour par la poste étaient inacceptables en raison des biais qu'ils étaient susceptibles d'introduire. La méthode de l'interview sur place offrait certains avantages, comme un faible taux de non-réponse et un faible taux de non-réponse à certaines questions, qui auraient permis d'améliorer la qualité des données recueillies, mais elle avait pour inconvénient d'être très coûteuse. De plus, nombre des plans d'échantillonnage utilisés pour réduire le coût des interviews sur place prévoient un échantillonnage à plusieurs degrés et sont fortement optimisés pour quelques variables. (Il aurait été trop coûteux de ne pas utiliser un des plans d'échantillonnage et une des bases de sondage existants.) Ces plans complexes rendent l'analyse des ensembles de données résultants difficile, tandis que l'optimisation se traduit par des effets de plan marqués pour certaines variables. Ces effets de plan font que ces plans d'échantillonnage conviennent moins bien aux enquêtes à objectifs multiples comme l'ESG. Par ailleurs, les enquêtes téléphoniques réalisées à Statistique Canada indiquaient que ces dernières permettaient d'obtenir des taux de réponse relativement élevés à un coût raisonnable. En outre, l'utilisation de méthodes de composition aléatoire (CA) constitue une façon efficace pour prélever des échantillons qui sont des échantillons aléatoires simples ou peu s'en faut.

Pour toutes ces raisons, l'ESG a utilisé les méthodes de composition aléatoire (CA) et d'interview téléphonique pour la majeure partie de son échantillon dans tous les cycles réalisés jusqu'à maintenant. Lorsqu'il s'est avéré nécessaire d'axer l'enquête sur certains groupes cibles, on a complété l'échantillon principal en prélevant des noms dans des bases de sondage sous forme de listes. À l'occasion du premier cycle de l'enquête, on a jugé qu'il fallait utiliser la technique de l'interview sur place pour nombre des interviews réalisées avec des personnes âgées.

### 3.3 Population cible

La population cible de l'ESG comprend toutes les personnes de 15 ans ou plus résidant au Canada à l'exception: i) des résidents des Territoires du Nord-Ouest et du Yukon et ii) des pensionnaires d'établissements institutionnels. Cette population cible est différente de celle de l'EPA, qui exclut en outre tous les habitants des réserves indiennes et les membres permanents des Forces armées canadiennes.

### 3.4 Population échantillonnée

Les méthodes d'échantillonnage utilisées pour l'ESG font que certains membres de la population cible sont exclus de l'échantillon. Au moment de la pondération, ces personnes sont implícitement traitées de la même façon que les membres de l'échantillon (manquantes au hasard) et les poids finals permettent d'établir des estimations pour l'ensemble de la population cible.



Chaque cycle de l'ESG comporte trois composantes:

- La composante de la thématique principale, dont chacun des sujets est étudié tous les cinq ans, vise à recueillir des données permettant de suivre l'évolution des conditions de vie et du niveau de bien-être.
- La composante de la thématique particulière, dont le sujet varie d'un cycle à l'autre, vise l'atteinte du deuxième objectif de l'enquête, soit fournir des renseignements sur des questions précises d'intérêt public présentant un intérêt particulier pour les ministères fédéraux ou les groupes de réflexion.

- La composante des données signalétiques. Ces données, qui sont recueillies à chaque cycle d'enquête, portent sur un ensemble de variables démographiques et socio-économiques de base qui permettent de définir les groupes de population afin de faciliter l'analyse des données thématiques principales et particulières.

Alors que les composantes des données signalétiques et de la thématique principale sont financées par Statistique Canada, les coûts relatifs à la composante de la thématique particulière sont récupérés auprès des paratrans.

La population cible de l'ESG comprend toutes les personnes âgées de 15 ans ou plus vivant dans les dix provinces du Canada, à l'exception des pensionnaires des établissements institutionnels. Il a été décidé de ne pas utiliser l'Enquête sur la population active comme véhicule pour l'ESG afin d'éviter d'imposer un fardeau de déclaration trop lourd aux personnes participant à l'EPA et de permettre à l'équipe de l'ESG d'utiliser des méthodes d'échantillonnage et de collecte ainsi que des distributions d'échantillonnage différentes de celles de l'EPA. La taille cible de l'échantillon prélevé pour chaque cycle, qui est de 10,000 personnes, constitue une solution de compromis permettant de satisfaire le mieux possible aux exigences en matière de précision des estimations, de respect du budget et de durée de l'interview. Les paratrans ont toutefois la possibilité d'élargir la taille de l'échantillon pour une population cible ou une région géographique donnée.

Le premier cycle de l'enquête, portant sur la santé, a été réalisé à la fin de 1985, et les autres cycles ont été ultérieurement mis en oeuvre à des intervalles d'environ un an. Le cinquième cycle, portant sur la famille, a été réalisé au début de 1990 et la collecte des données relatives au sixième cycle a débuté en janvier 1991.

Avant d'étudier de façon plus approfondie les thèmes et sujets de recherche abordés dans le cadre de chacun des cycles de l'enquête, nous allons exposer plus en détails les méthodes d'enquête utilisées.

### 3. MÉTHODOLOGIE

#### 3.1 Exigences et contraintes

Voici les principales exigences méthodologiques auxquelles doit satisfaire l'ESG: i) elle doit permettre la réalisation d'une analyse approfondie de la population canadienne d'âge adulte à l'échelle nationale et d'une analyse un peu moins détaillée à l'échelle régionale (cette exigence a des implications tant sur le plan de la taille de l'échantillon que sur celui de la quantité de données recueillies auprès de chaque répondant); ii) elle doit être réalisée à un coût acceptable; iii) elle doit utiliser un plan d'échantillonnage qui convienne à une enquête à objectifs multiples; enfin, iv) elle doit permettre l'élaboration de fichiers de microdonnées à grande diffusion pouvant, sans trop de difficultés, être utilisés à des fins d'analyse par les chercheurs de l'extérieur.

Ces exigences ont toutes une incidence sur le choix du mode de collecte des données, du plan d'échantillonnage et de la taille de l'échantillon. Toutefois, le choix du plan d'échantillonnage a été déterminé surtout par les deux dernières, tandis que celui du mode de collecte et de la taille de l'échantillon l'a été dans une large mesure par le plan d'échantillonnage et les deux premières.

ne représentait qu'environ 10% du PIB, elles en constituaient environ 30% au début des années 80. Cette augmentation du nombre de programmes sociaux s'est accompagnée d'un accroissement de la demande de données et de renseignements permettant de suivre et d'analyser les tendances sociales ainsi que d'une intensification de l'utilisation de ces données et renseignements. Au fil des ans, Statistique Canada a dû élargir la portée de son programme de la statistique sociale pour répondre à ces besoins de plus en plus pressants. Néanmoins, l'utilisation plus étendue qu'on a faite des données disponibles au cours des dernières années a révélé l'existence de vastes domaines lacunaires où les données pertinentes étaient trop réduites et limitées pour permettre de planifier efficacement les programmes, les produits et les services ou de déterminer à quels projets il convenait d'affecter les ressources.

Au début des années 80, un des points faibles du programme tenait au fait que la plupart des statistiques sociales autres que celles portant sur le marché du travail et sur le revenu étaient élaborées à partir de dossiers administratifs ou d'enquêtes réalisées auprès des établissements institutionnels. Ces sources permettaient de recueillir des données limitées sur la population venant en contact avec les établissements institutionnels du secteur social, mais aucune donnée sur les besoins en matière de programmes sociaux ni sur l'incidence de ces programmes sur la population en général. La seule façon de recueillir de telles données consistait à réaliser une enquête auprès de cette population.

Même si nombre d'arguments militaient en faveur de la réalisation à intervalles réguliers d'enquêtes de grande envergure portant sur divers sujets (comme la santé, l'éducation, les victimes d'actes criminels), on ne disposait pas des ressources nécessaires pour mettre sur pied un programme d'une telle envergure. Statistique Canada a plutôt choisi de réaliser une enquête sociale générale annuelle beaucoup plus modeste permettant d'étudier les principaux sujets d'importance au cours d'une période de cinq ans, ainsi que servant, à long terme, d'outil pour suivre l'évolution de la société et, à court terme, d'outil de collecte de données limitées sur des questions présentant un intérêt sur le plan de la politique sociale. Le budget annuel total de l'ESG s'établissait initialement à environ un million de dollars (CAN) et le programme était financé au moyen d'une réaffectation interne des ressources dégagées par Statistique Canada à la suite d'un accroissement de l'efficacité de l'Enquête sur la population active.

#### L'ESG a deux objectifs principaux:

- recueillir à intervalles réguliers des données sur un large éventail de tendances sociales afin de suivre l'évolution de la société en ce qui concerne les conditions de vie et le bien-être des Canadiens;
- fournir des renseignements sur des questions précises d'actualité ayant trait à la politique sociale.

En vue d'atteindre ces objectifs, on a décidé de faire de l'ESG une enquête annuelle. Étant donné le large éventail de questions sociales sur lesquelles il faut recueillir des données, l'ESG comporte cinq cycles d'enquêtes portant chacun sur un sujet différent. Les données relatives à chacun des sujets traités sont donc recueillies tous les cinq ans. Les divers cycles de l'enquête portent respectivement sur:

1. La santé;
2. L'emploi du temps;
3. Les risques personnels (accidents et actes criminels);
4. Les études et le travail;
5. La famille et les amis.

Lors de la planification du contenu, on s'était également fixé comme objectif d'inclure des questions permettant d'obtenir des indicateurs de la qualité de vie, questions portant par exemple sur le degré de satisfaction, les attitudes, les perceptions ou les croyances.

# L'Enquête sociale générale: bilan des cinq premières années

D.A. NORRIS et D.G. PATON<sup>1</sup>

## RÉSUMÉ

L'Enquête sociale générale canadienne est une enquête annuelle ayant pour objet de recueillir des données sur les caractéristiques démographiques et sociales des Canadiens. Le présent article donne un aperçu général de l'enquête fondé sur l'expérience des cinq premiers cycles. Nous y examinons les objectifs du programme, les méthodes utilisées, les thèmes et les questions étudiés, les produits élaborés ainsi que les perspectives d'avenir.

**MOTS CLÉS:** Enquêtes sociales; enquêtes téléphoniques; composition allochtone; enquêtes sur l'emploi du temps; enquêtes sur la santé.

## 1. INTRODUCTION

Le programme de la statistique sociale de Statistique Canada a pour objet de fournir des renseignements sur les caractéristiques démographiques et sociales des Canadiens ainsi que sur leurs conditions de vie. Les produits conçus dans le cadre du programme servent de base à l'élaboration des politiques relatives à nombre de questions sociales capitales.

La pierre angulaire du programme de la statistique sociale est le recensement de la population. Tenu tous les cinq ans, ce dernier permet de recueillir des données repères sur l'évolution démographique, sociale et économique de la population et constitue la base pour la réalisation d'enquêtes-échantillon ultérieures. En sus du recensement, le programme prévoit la mise en oeuvre d'enquêtes permanentes et d'autres programmes statistiques, dont un bon nombre utilisent des sources de données administratives, portant sur la santé, l'éducation, la culture, la justice, les finances publiques, l'emploi et le chômage, le revenu et les dépenses ainsi que la démographie.

Bien que les enquêtes-ménages constituent depuis longtemps un volet important du programme de la statistique sociale, les enquêtes périodiques ont traditionnellement porté davantage sur les questions relatives au marché du travail et au revenu et aucune enquête périodique permanente n'a été réalisée dans des domaines comme la santé, l'éducation, la justice ou la culture. Afin de combler partiellement ces lacunes, Statistique Canada a mis sur pied en 1985 une Enquête sociale générale (ESG).

Le présent article se propose d'exposer la nature et la portée de l'ESG, puis d'indiquer comment elle a évolué au cours des cinq dernières années. Vous y trouverez une description des méthodes utilisées et des sujets traités dans le cadre des cinq cycles de l'enquête, suivie d'une brève discussion des pistes qu'il conviendrait d'explorer dans le futur.

## 2. OBJECTIFS ET STRUCTURE DE L'ESG

Nous avons assisté, au cours de la période allant de 1930 à 1980, à une croissance rapide du nombre et de l'envergure des programmes sociaux mis en oeuvre au Canada. Alors qu'au début des années 30 l'ensemble des dépenses gouvernementales au titre des programmes sociaux

<sup>1</sup> D.A. Norris et D.G. Paton, Statistique Canada, Ottawa (Ontario) Canada, K1A 0T6.



REMERCIEMENTS

Cette étude a été rendue possible en partie grâce au contrat de coopération n° 58-3AEU-9-80040 passé avec le Département de l'agriculture des États-Unis. Les commentateurs qui y sont exprimés engagent la responsabilité des auteurs seuls. Ceux-ci tiennent à remercier Gary Keough et Leland Brown, du Département de l'agriculture des E.-U., de même que Ronald Sadler, Melvin Perrott, Eldon Thiesse et M.E. Johnson, du Département de l'agriculture du Kansas, pour les avoir aidés dans ce projet. Ils remercient également l'arbitre et le rédacteur associé qui ont bien voulu exprimer leurs commentaires sur une version antérieure de l'article.

BIBLIOGRAPHIE

BASS, J., GUINN, B., KLUGH, B., RUCKMAN, C., THORSON, J., et WALDROP, J. (1989). Report of the Task Group for Review and Recommendations on County Estimates. USDA National Agricultural Statistics Service, Washington, D.C.

FAY, R.E., et HERRIOT, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

KANSAS AGRICULTURAL STATISTICS. (1988). Kansas Farm Facts, préparé par le Statistical Division of the Kansas Department of Agriculture avec la coopération du National Agricultural Statistics Service of the U. S. Department of Agriculture, Topeka, Kansas.

PLATEK, R., RAO, J.N.K., SÄRNDALE, C.E., et SINGH M.P. (Eds.) (1987). *Small Area Statistics*. New York: John Wiley & Sons.

SÄRNDALE, C.E., et HIDIROGLOU, M.A. (1989). Small domain estimation: A conditional analysis. *Journal of the American Statistical Association*, 84, 266-275.

La somme des estimations régionales initiales est voisine du total pour la population simulée. Les facteurs de pondération constants, ci, ont donc tous une valeur approximative de un. Enfin, comme les deux méthodes donnent des résultats similaires, il n'y a aucune raison d'utiliser la plus compliquée des deux; nous pouvons donc utiliser celle avec facteur de pondération constant.

## 7. CONCLUSIONS ET SUJETS DE RECHERCHE POUR L'AVENIR

Nous avons montré que l'on pouvait obtenir des estimations acceptables de la production de blé par comté à l'aide d'un modèle de régression. Les variables explicatives du modèle que nous avons choisi sont la superficie ensemencée (en acres), les indicateurs de district et les termes d'interaction. Le modèle de régression proposé ne nécessite pas un échantillon aléatoire de fermes et, par surcroît, il permet d'estimer la variance d'estimations régionales. Il est possible de pondérer les estimations tirées du modèle de régression de manière à faire concorder leur somme avec la production totale estimée pour l'Etat; cette pondération peut se faire au moyen d'un facteur constant car nous avons vu que cette méthode ne donne pas des résultats très différents de ceux obtenus avec l'autre méthode, plus complexe.

L'estimation de la production agricole par comté est un sujet qui doit faire encore l'objet de nombreuses recherches. Par exemple, les estimations régionales tirées de l'étude de simulation nous portent à croire que la présence (ou l'absence) des grandes exploitations agricoles (embalvures d'une superficie de 1,000 acres ou plus) dans l'échantillon d'un district pourrait influencer largement sur les estimations des comtés de ce district, à plus forte raison pour les districts qui comptent peu de grandes exploitations. Comme les grosses fermes représentent le plus souvent une proportion appréciable de la production agricole, elles mériteraient de faire l'objet d'une opération indépendante dans le calcul d'estimations régionales. Par ailleurs, les Etats devraient envisager de modifier leur plan d'échantillonnage de sorte que les grandes exploitations fassent partie de l'échantillon à coup sûr.

La recherche doit aussi tenter de déterminer si le modèle de régression que nous avons défini pour le blé peut s'appliquer à d'autres cultures. En particulier, il serait intéressant de savoir si ce genre de modèle peut être utilisé pour les cultures spéciales, au sujet desquelles il existe beaucoup moins de données. De plus, nous découvrirons peut-être que la similitude observée entre le chiffre de production pour l'Etat et la somme des estimations régionales – pour les données réelles pour les données des échantillons simulés – est propre à la production de blé et n'est pas observable pour toutes les cultures. Nous devrions aussi chercher à savoir si, pour les autres cultures que le blé, on doit utiliser une autre méthode de pondération que celle avec facteur de pondération constant.

Nous avons choisi d'amorcer l'étude de l'estimation pour comtés en examinant des méthodes d'estimation de la production. Les recherches à venir devront porter aussi sur l'estimation de la superficie totale ensemencée pour diverses cultures. Pour les besoins de notre étude, nous avons puisé dans la base de données du recensement de l'agriculture de 1987 pour connaître le nombre de fermes et la superficie ensemencée en blé dans chaque comté. Or, le recensement de l'agriculture n'est fait qu'à tous les cinq ans. Dans l'intervalle, il faut estimer la variation du nombre de fermes et de la superficie ensemencée à l'aide de données d'échantillon. Ces variations devraient normalement être modestes en ce qui a trait aux grandes cultures comme le blé au Kansas; toutefois, pour ce qui regarde les cultures moins courantes, il sera vraisemblablement plus difficile d'estimer ces variations.

Enfin, il ne sera probablement plus nécessaire de respecter la condition que nous avons posée au début – à savoir l'utilisation d'un estimateur comportant peu de calculs – et qui nous a amenés à proposer un estimateur par régression, car les divers départements de l'agriculture aux Etats-Unis seront bientôt rattachés à un grand réseau informatique national. Par conséquent, nos futures études sur l'estimation de la production agricole par comté s'intéresseront à des estimateurs qui comportent plus de calculs.

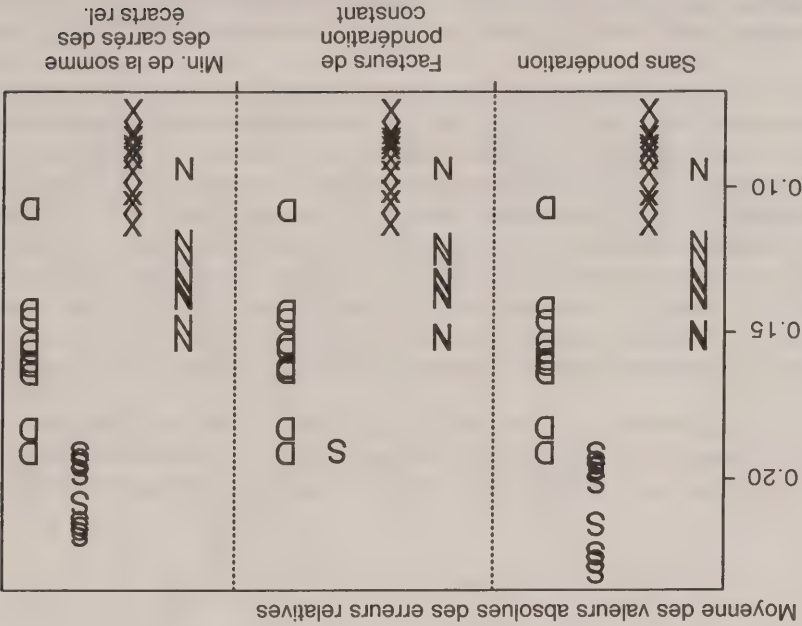
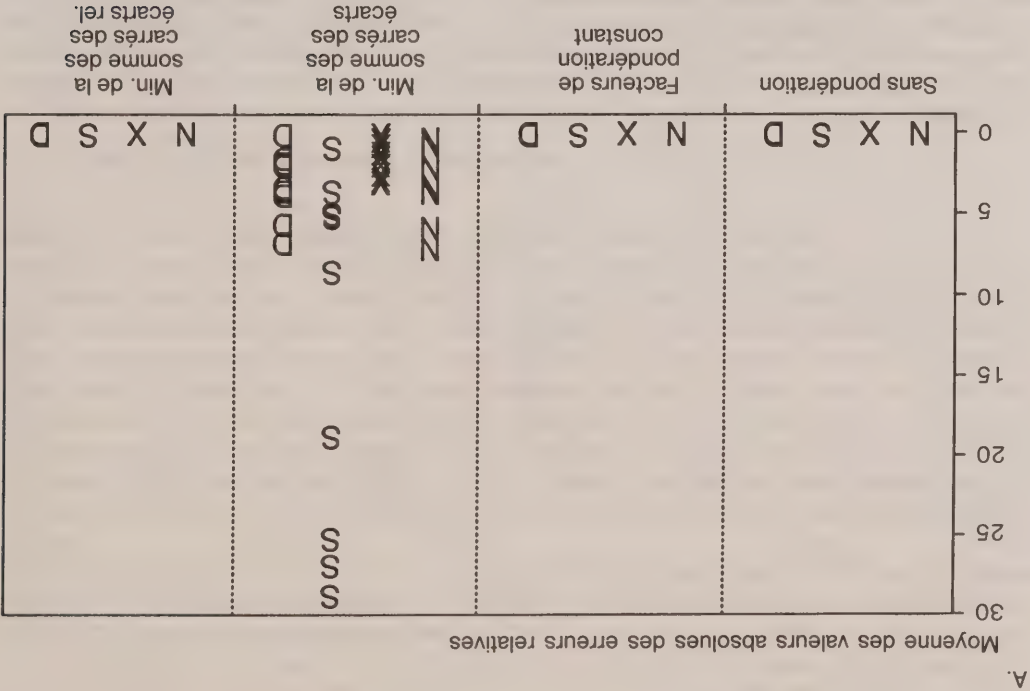


Figure 2. Comparaison des estimateurs et des méthodes de pondération

Nota: Les données proviennent des échantillons simulés.  
N = estimateur par régression – modèle 1 (sans terme d'interaction);  
X = estimateur par régression – modèle 2 (avec termes d'interaction);  
S = estimateur synthétique;  
D = estimateur direct.



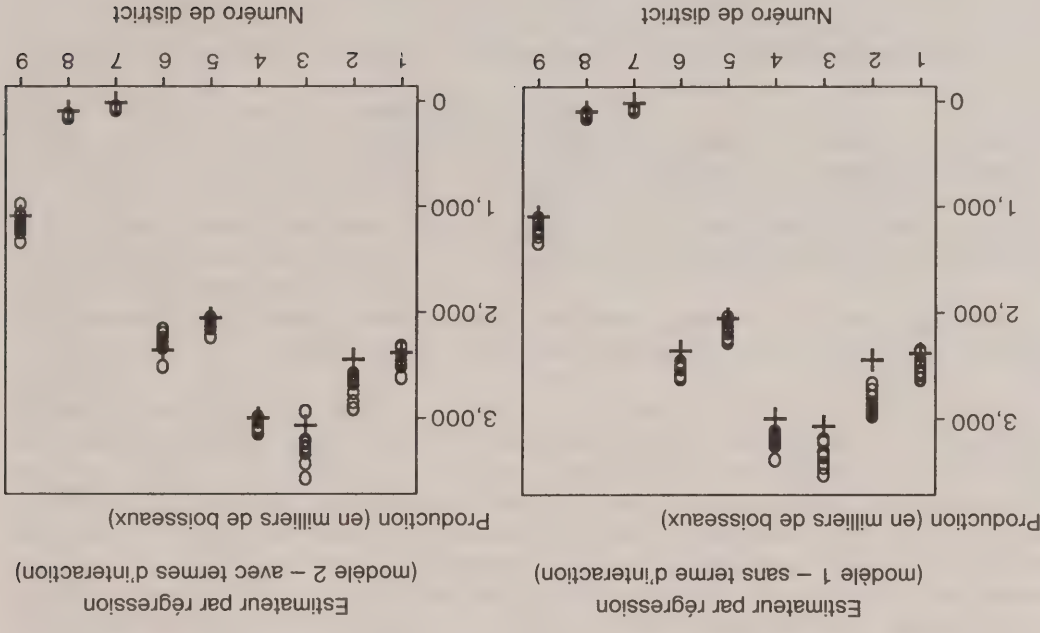
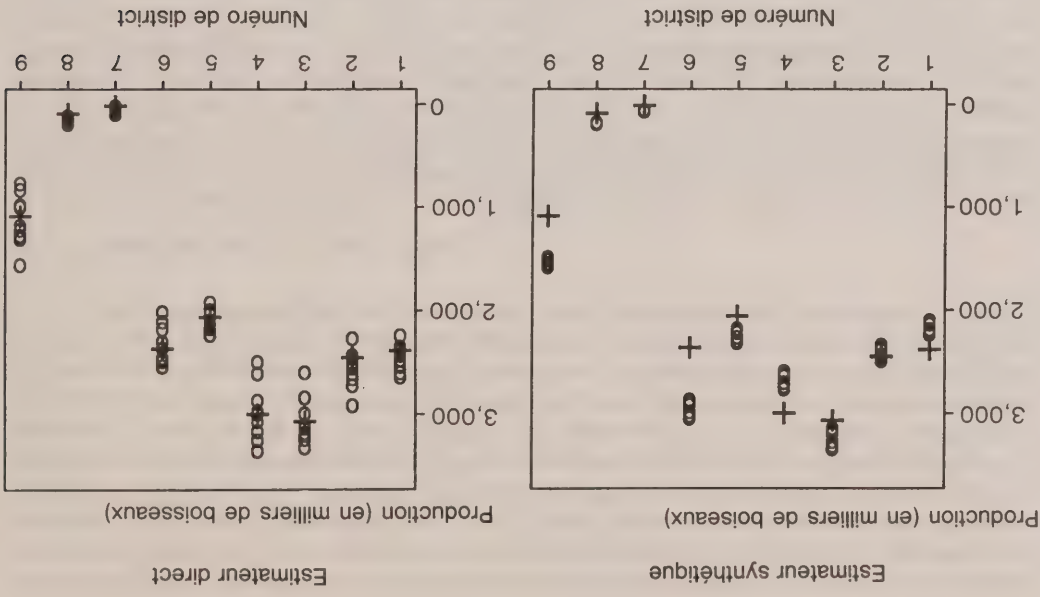
des 105 comtes. Nous avons ensuite comparé ces estimations aux niveaux de production "réels" établis à l'aide de la population simulée. Cette comparaison nous permet de mesurer le biais et la variabilité des estimations pour chaque comté. Dans la figure 1, nous reproduisons, pour chaque estimateur étudié et les neuf comtes choisis au hasard à raison d'un par district (évoqués dans la section 5), les valeurs calculées pour chacun des dix échantillons ainsi que la production (réelle). Comme prévu, les estimations synthétiques sont entachées d'un biais considérable. En effet, l'intervalle des estimations issues des dix échantillons simulés ne comprend la valeur réelle de la population que dans le cas du district n° 2. L'intervalle des valeurs calculées par l'estimateur direct est plus large que celui des estimations synthétiques pour chacun des neuf districts mais il renferme la valeur réelle de la population dans tous les cas. Les intervalles des estimations calculées au moyen de l'un ou l'autre des modèles de régression semblent moins étendus que ceux des estimations directes. Dans le cas du modèle 1, on note un certain biais dans les estimations pour environ la moitié des districts. Les estimations calculées selon le modèle 2 paraissent moins biaisées. Les résultats de la comparaison nous amènent à opter pour le modèle 2, c'est-à-dire le modèle de régression avec termes d'interaction.

### 6.4 Comparaison des méthodes de pondération

En nous servant des trois méthodes de pondération décrites dans la section 4, nous avons ensuite pondéré les quatre mêmes séries d'estimations régionales tirées des échantillons simulés de manière à les faire concorder avec la production totale de l'Etat estimée à l'aide de la population simulée. Nous avons comparé les estimations ainsi obtenues aux niveaux de production réels établis pour la population simulée. A cette fin, nous sommes servis de la moyenne des valeurs absolues des erreurs relatives, qui est définie:

$$(1/I) \sum_{i=1}^I \left| (X_{i+} - Y_{i+}) / Y_{i+} \right|.$$

La figure 2 donne les valeurs de cette moyenne tirées des dix échantillons simulés pour les 105 comtes. Nous y comparons les quatre estimateurs dans le cas où il y a absence de pondération et dans trois autres cas qui correspondent à l'application de chacune des trois méthodes de pondération. Nous voyons d'après la figure 2A que la méthode qui consiste à minimiser la somme des carrés des écarts donne des estimations finales très médiocres; la moyenne des valeurs absolues des écarts relatifs entre les estimations finales et les niveaux de production des comtes calculés à l'aide de la population simulée est très élevée par rapport aux moyennes observées dans le cas des autres méthodes de pondération. Cela s'explique par la grande différence qui peut exister parfois entre des niveaux de production dans deux comtes. Comme la méthode en question consiste à minimiser la somme des carrés des écarts entre les estimations initiales et les estimations finales, il se pourrait que l'opération modifie considérablement l'estimation initiale pour un comté si cette estimation est très peu élevée. Or, de telles variations semblent injustifiables et c'est pourquoi nous n'insisterons pas davantage sur cette méthode de pondération. La figure 2B, qui est une version raffinée de la figure 2A, permet une comparaison plus poussée des quatre estimateurs dans le cas de l'absence de pondération et dans le cas de l'application des deux autres méthodes. Nous constatons que le modèle de régression avec termes d'interaction est, parmi les quatre estimateurs étudiés, celui pour lequel l'erreur est généralement la moins élevée. Cela confirme le choix que nous avions fait dans la sous-section précédente. En outre, la figure 2B donne à penser qu'il y a peu de différence entre les estimations initiales non pondérées et les estimations pondérées selon l'une ou l'autre des deux méthodes. De fait,



Nota: Les estimations concernent un comté tiré au hasard dans chacun des districts.  
o = estimation calculée pour un des dix échantillons simulés;  
+ = valeur réelle tirée de la production simulée.

Figure 1. Comparaison d'estimateurs pour neuf comtés

Comme l'estimation directe calculée pour un comté repose uniquement sur les données d'échantillon recueillies à l'intérieur de ce comté, la variance de cette estimation sera relativement élevée mais son biais devrait être moindre que celui d'une estimation synthétique. L'échantillon doit comprendre au moins une ferme d'un comté donné si l'on veut calculer une estimation pour ce comté, et il doit comprendre au moins deux fermes d'un même comté si l'on veut estimer la variance. Dans la base de données de 1987 sur les menus grains, trois comtés n'étaient pas du tout représentés dans l'échantillon des fermes à blé et trois autres l'étaient par une seule ferme. Bien que nous comparions "nos" estimateurs par régression aux estimateurs synthétique et direct, il convient de préciser que ces deux derniers n'admettent que des données provenant d'un échantillon probabiliste, ce qui n'est pas le cas des données du Kansas.

6.2 Population simulée et échantillons

Nous avons tout d'abord simulé une population de fermes à blé en générant des niveaux de production pour les 22,300 fermes de la base de données sur les emblavures qui comprenaient des terres ensemençées en blé. Comme les niveaux de production semblent varier selon le district et la taille de l'exploitation agricole (voir tableau 3), nous avons généré des ratios "nombre de boisseaux par unité de superficie ensemençée" (bu/pa) à partir de 37 distributions différentes. Celles-ci étaient fondées sur les ratios bu/pa tirés de la base de données sur les menus grains. (Notons qu'en ce concerne les districts de l'est du Kansas – districts 7, 8 et 9 en l'occurrence, plusieurs catégories de taille étaient peu représentées dans l'échantillon ou ne l'étaient pas du tout. Nous avons donc regroupé ces catégories de la manière indiquée dans le tableau 3 afin de simuler des ratios bu/pa.) Nous avons produit un histogramme des ratios empiriques de la base de données sur les menus grains pour chaque district en fonction de cinq catégories de taille: de 0 à 99 acres ensemençées en blé, de 100 à 249, de 250 à 499, de 500 à 999 et 1,000 acres ou plus. Comme ces histogrammes avaient plus ou moins une forme en cloche, nous avons choisi de décrire la distribution des ratios bu/pa au moyen de la distribution normale. Pour moyenne et pour variance de la distribution normale, nous avons utilisé la moyenne et la variance empiriques des ratios bu/pa pour les fermes à blé de la base de données sur les menus grains réparties dans les 37 classes formées par la combinaison des districts et des catégories de taille.

Après avoir tiré de la distribution normale appropriée le ratio bu/pa pour chaque ferme, nous avons calculé la quantité de blé produite sur chaque ferme en multipliant le ratio simulé par la superficie déclarée des emblavures. Par la suite, nous avons tiré dix échantillons de cette population simulée. Comme aucun plan d'échantillonnage particulier n'avait été défini, chaque ferme dans les 37 classes a été prélevée avec une probabilité égale au rapport empirique entre le nombre de fermes appartenant à la base de données sur les menus grains et le nombre de fermes de la même classe appartenant à la base de données sur les emblavures. Ainsi, les fermes de la classe C, par exemple, ont été échantillonnées avec une probabilité égale à

$$\frac{\text{Nombre de fermes de la classe C dans la base de données sur les menus grains}}{\text{Nombre de fermes de la classe C dans la base de données sur les emblavures}}$$

Par un tel plan, nous cherchions à obtenir des échantillons simulés qui se rapprocheraient le plus possible des échantillons réels, même si nous ne connaissons pas les probabilités de sélection pour les échantillons réels.

6.3 Comparaison des quatre estimateurs pour petites régions

Nous nous sommes servis des quatre estimateurs qui font l'objet d'une comparaison (c'est-à-dire les deux estimateurs par régression, l'estimateur synthétique et l'estimateur direct) et des dix échantillons simulés précédemment pour calculer un niveau de production pour chacun



Särndal et Hidiroglou (1989) pour une analyse d'estimateurs courants pour petites régions, y compris les estimateurs synthétique et direct.) On obtient des estimations synthétiques en répartissant la production totale de blé pour l'Etat entre les comités au prorata de la superficie enssemencée en blé dans chaque comité. Quant à l'estimation directe, la méthode se limite à utiliser les fermes échantillonnées d'un comité pour estimer la production de blé dans ce comité. Les estimations synthétiques devraient normalement être fortement biaisées puisque les pratiques agricoles et les conditions atmosphériques ne sont pas nécessairement les mêmes dans chaque région d'un Etat alors que selon l'estimateur synthétique, chaque comité est représentatif de l'Etat. En revanche, la variance des estimations synthétiques sera relativement faible car ce genre d'estimations sont établies à l'aide de toutes les données pour l'Etat.

Tableau 3

Nombre de fermes et niveau de production selon le district et la superficie enssemencée

District	Superficie enssemencée (acres)				
	0-99	100-249	250-499	500-999	≥ 1,000
1	$M_i^*$ 354 $m_i^*$ 27 bu/pa*	638 45 37.18	531 51 37.76	302 40 39.21	85 9 38.68
2	$M_i$ 266 $m_i$ 27 bu/pa	550 49 33.62	572 47 36.78	377 55 39.09	161 33 34.85
3	$M_i$ 264 $m_i$ 31 bu/pa	549 80 32.84	610 76 35.03	537 98 36.79	264 61 33.13
4	$M_i$ 956 $m_i$ 62 bu/pa	939 37 36.91	626 23 39.70	271 21 39.87	50 7 39.41
5	$M_i$ 1,236 $m_i$ 92 bu/pa	1,529 93 32.25	912 51 31.69	350 26 36.85	54 3 33.65
6	$M_i$ 1,181 $m_i$ 96 bu/pa	1,427 96 26.88	1,160 81 28.78	793 55 27.87	249 20 26.72
7	$M_i$ 957 $m_i$ 62 bu/pa	242 5 40.81**	67 2 40.81**	9 0 40.81**	3 0 40.81**
8	$M_i$ 1,126 $m_i$ 56 bu/pa	251 11 11.48**	52 2 11.48**	9 0 11.48**	1 0 11.48**
9	$M_i$ 1,122 $m_i$ 47 bu/pa	431 19 23.87	166 7 27.63**	59 3 27.63**	12 1 27.63**

\*  $M_i$  est le nombre de fermes dans la base de données sur les emblavures,  $m_i$  est le nombre de fermes dans la base de données sur la production et bu/pa est le rapport entre le nombre de boisseaux produits et la superficie enssemencée (en acres).  
\*\* Les cellules pour ce district ont été groupées en vue de calculer le ratio bu/pa.

Tableau 2

Estimations par régression pour neuf comtés du Kansas

Production estimée de blé (en milliers de boisseaux)		
District	Comté	
Modèle 1 (sans terme d'interaction)		
Modèle 2 (avec termes d'interaction)		
1	Decatur	4,944 (180)
	Trego	4,378 (174)
	Hodgeman	4,808 (123)
	Jewell	5,555 (275)
	Marion	5,144 (313)
	Comanche	2,615 (59)
	Leavenworth	231 (53)
	Shawnee	232 (106)
	Butler	2,374 (331)
2	Decatur	4,778 (179)
	Trego	4,229 (188)
	Hodgeman	4,908 (125)
	Jewell	5,550 (269)
	Marion	4,931 (315)
	Comanche	2,480 (63)
	Leavenworth	262 (61)
	Shawnee	226 (104)
	Butler	2,272 (338)

Nota: Les erreurs types figurent entre parenthèses au-dessous de chaque estimation.

Nous nous sommes servis des deux modèles proposés pour établir des estimations pour les 105 comtés du Kansas. Le tableau 2 donne les estimations non pondérées et les erreurs types correspondantes calculées pour neuf comtés, soit un par district. Nous avons tiré au hasard un comté dans chaque district de sorte que toutes les régions de l'Etat soient représentées. En examinant les données du tableau 2, on s'aperçoit que l'erreur type estimée pour le comté de Shawnee est anormale. La variance d'une estimation régionale (ou de comté) dépend du nombre de fermes dans le comté, de la superficie totale ensemencée en blé dans le comté et du nombre de fermes de l'échantillon qui appartiennent au district dans lequel se trouve le comté. Le district no 8, qui comprend Shawnee, compte relativement peu de fermes dans la base de données sur les menus grains. Par ailleurs, Shawnee a un nombre raisonnable de fermes où l'on cultive du blé mais les superficies ensemencées sont modestes. Ces trois facteurs expliquent que l'erreur type des estimations calculées pour le comté de Shawnee soit assez élevée.

Les estimations du tableau 2 se rapprochent assez des estimations publiées (Kansas Agricultural Statistics 1988). Bien qu'il soit réconfortant de voir que nos estimations ne s'écartent pas véritablement de celles de l'Etat du Kansas, il n'y a rien théoriquement qui justifie de considérer les estimations du Kansas comme la norme. Nous avons donc exécuté une étude de simulation afin de mieux évaluer nos estimateurs. Cette étude est décrite dans la section qui suit.

6. ETUDE DE SIMULATION

6.1 Estimateurs comparés

L'étude de simulation a permis de comparer les estimations obtenues au moyen des deux modèles de régression proposés et celles produites par deux estimateurs courants pour petites régions: l'estimateur synthétique et l'estimateur direct. (Voir, par exemple, la section 2 de

Tableau 1

Ajustement des modèles de régression aux données réelles

Modèles ajustés		$R^2$		$\sqrt{cme}$
Modèle 1	Boisseaux = - 811 + 32(P1a) + 3,248I <sub>1</sub> + 3,088I <sub>2</sub> + 2,190I <sub>3</sub>	85	5,945	+ 2,526I <sub>4</sub> + 1,241I <sub>5</sub> - 562I <sub>6</sub> + 1,047I <sub>7</sub> + 399I <sub>8</sub>
Modèle 2	Boisseaux = - 281 + 28(P1a) + 138I <sub>1</sub> + 1,861I <sub>2</sub> + 2,328I <sub>3</sub>	86	5,818	+ 329I <sub>4</sub> - 359I <sub>5</sub> - 334I <sub>6</sub> - 42I <sub>7</sub> + 500I <sub>8</sub>
				+ 11(P1a)I <sub>1</sub> + 5(P1a)I <sub>2</sub> + 3(P1a)I <sub>3</sub> + 11(P1a)I <sub>4</sub>
				+ 9(P1a)I <sub>5</sub> - 0,2(P1a)I <sub>6</sub> + 15(P1a)I <sub>7</sub> - 7(P1a)I <sub>8</sub>

Nota: P1a désigne la superficie ensemencée (en acres); I<sub>i</sub> est la variable indicatrice pour le i-ème district.

Les variables les plus importantes pour notre modèle de régression étaient la superficie ensemencée en blé et un indicateur de l'emplacement de la ferme sur le territoire de l'Etat. La variable fondée sur les estimations par comté de l'année précédente ne semblait pas un bon prédicteur de la quantité de blé produit sur une ferme dans l'année courante. Nous aurions pu aussi choisir d'autres variables comme la superficie des emblavures irriguées mais il est difficile d'en connaître la valeur exacte au niveau des comtés. C'est pourquoi nous avons choisi de faire de la superficie ensemencée en blé la seule variable explicative contenue de notre modèle. Nous n'étions pas obligés d'inclure tous les indicateurs de district dans le modèle; certains districts étaient similaires et auraient donc pu être regroupés. Si nous avons décidé de tenir compte de tous les indicateurs, c'est que les groupes de districts risquaient de changer d'une année à l'autre ou pouvaient être différents dans le cas d'autres cultures que le blé.

Nous avons choisi de concentrer notre attention sur deux modèles de régression: le modèle 1, qui comprend la superficie ensemencée en blé et les indicateurs de district, et le modèle 2, qui comprend les mêmes variables plus les termes d'interaction qui mettent en rapport la superficie ensemencée et les variables indicatrices. Ces modèles et les mesures d'ajustement correspondantes figurent dans le tableau 1. Malgré la faible différence entre les racines du carré moyen des erreurs ( $\sqrt{cme}$ ) des deux modèles, nous étions d'avis que cette différence pourrait s'amplifier si les modèles devaient servir à estimer la production agricole pour tout l'Etat. C'est pourquoi nous avons cru utile de comparer des estimations tirées des deux modèles.

Pour nous assurer que ces modèles de régression n'étaient pas simplement la conséquence d'une irrégularité dans la base de données de 1987 sur les menus grains, nous avons repris le même ensemble de variables explicatives et avons tenté de construire des modèles de régression acceptables à l'aide des données de 1988. L'ajustement des modèles 1 et 2 aux données de 1988 est comparable à l'ajustement des mêmes modèles aux données de 1987 et nous n'avons trouvé aucun autre modèle qui pouvait produire un meilleur ajustement pour les données de 1988. Les valeurs estimées du paramètre correspondant à la superficie ensemencée en blé étaient assez semblables d'une année à l'autre, alors qu'il y avait des différences notables entre les paramètres pour les indicateurs de district. Nous croyons que ces indicateurs traduisent l'effet des conditions atmosphériques et des différences de pratiques agricoles dans les diverses parties de l'Etat. Par exemple, on irrigue beaucoup plus les terres dans l'ouest et le centre du Kansas que dans l'est. Les pratiques agricoles sont peu susceptibles de changer considérablement d'une année à l'autre mais il en va tout autrement des conditions atmosphériques. C'est pourquoi il est raisonnable de penser que la contribution de l'indicateur de district dans la prédiction de la production de blé pourra varier sensiblement d'une année à l'autre.



Notons que l'on obtient l'estimation pondérée  $\tilde{Y}_{i+}$ , en ajoutant à l'estimation initiale  $Y_{i+}$ , un rapport qui a pour numérateur l'écart entre l'estimation de la production totale de l'Etat calculée par le DABU et la somme des estimations initiales par comté, et pour dénominateur une expression fondée sur la moyenne harmonique des estimations initiales. Bien qu'en théorie, certaines des estimations pondérées peuvent avoir une valeur négative, cela risque peu de se produire dans la réalité puisque les agriculteurs ont tendance à sous-estimer leur production dans les rapports qu'ils fournissent. Si jamais la somme des estimations initiales par comté était supérieure à l'estimation calculée par le DABU pour l'Etat, l'estimation pondérée se rapportant à un comté pour lequel l'estimation initiale est une valeur peu élevée pourrait être négative.

3) Choisir les  $c_i$  de manière à minimiser la somme des carrés des écarts relatifs entre  $\tilde{Y}_{i+}$  et  $Y_{i+}$

Si nous voulons choisir les  $c_i$  de manière à minimiser la somme des carrés des écarts relatifs entre  $\tilde{Y}_{i+}$  et  $Y_{i+}$ , pourvu que  $\sum_{i=1}^I c_i Y_{i+} = Y$ , nous devons minimiser  $\sum_{i=1}^I [(Y_{i+} - \tilde{Y}_{i+}) / Y_{i+}]^2 = \sum_{i=1}^I (c_i - 1)^2$  par rapport à  $c_i$  au moyen d'un multiplicateur de Lagrange afin d'imposer la condition voulue. De cette manière, nous observons que

$$c_i = 1 + \left[ Y_{i+} \left( Y - \sum_{i=1}^I Y_{i+} \right) / \sum_{i=1}^I Y_{i+}^2 \right]$$

et

$$\tilde{Y}_{i+} = Y_{i+} + \left[ Y_{i+}^2 \left( Y - \sum_{i=1}^I Y_{i+} \right) / \sum_{i=1}^I Y_{i+}^2 \right].$$

Là encore, on obtient l'estimation pondérée  $\tilde{Y}_{i+}$ , en ajoutant à l'estimation initiale  $Y_{i+}$ , un rapport qui a pour numérateur une expression axée sur l'écart entre l'estimation de la production totale de l'Etat calculée par le DABU et la somme des estimations initiales par comté, et pour dénominateur la somme des carrés des estimations initiales. Comme dans le scénario précédent, l'estimation pondérée peut avoir une valeur négative mais cela risque peu de se produire dans la pratique.

Il convient de souligner que dans l'expression à minimiser, nous avons choisi d'exprimer l'écart  $Y_{i+} - \tilde{Y}_{i+}$  par rapport à  $Y_{i+}$  plutôt que par rapport à  $\tilde{Y}_{i+}$ . Comme nous visions à construire des estimateurs qui comportent peu de calculs, nous avons choisi de considérer le rapport entre l'écart  $Y_{i+} - \tilde{Y}_{i+}$  et  $Y_{i+}$ .

Dans la section qui suit, nous examinons l'effet de ces trois méthodes de pondération sur les estimations régionales de la production de blé.

5. COMPARAISON DES ESTIMATIONS DE LA PRODUCTION DE BLE

Comme nous l'avons vu dans la section 3, nous nous sommes servis d'un modèle de régression linéaire pour décrire la relation entre la production de blé (mesurée en boisseaux) et certaines variables explicatives en ce qui a trait aux fermes de la base de données de 1987 sur les menus grains. Les variables explicatives que nous pouvions considérer étaient les suivantes: superficie ensemencée en blé, superficie moissonnée, production prévue (suivant les estimations par comté de 1986), superficie des emblavures irriguées, superficie des emblavures non irriguées, indicateurs du district où est située la ferme, indicateurs de la région de l'Etat (est, centre, ouest) et termes d'interaction.

Bien que  $Z$  soit inconnue, le produit  $A^T Z$  est une matrice connue qui est formée uniquement des  $N_i$  (nombre de fermes dans un comté) et des  $X_{i+k}$  (valeur totale d'une variable explicative pour un comté). Par conséquent, si nous utilisons le carré moyen des erreurs (c.m.e.) comme estimation de  $\sigma^2$ , nous pouvons obtenir des valeurs estimées de la variance des estimations par comté, ce qui ne s'était jamais fait jusqu'à maintenant.

L'estimateur par régression que nous venons de décrire satisfait aux conditions que nous avions fixées au départ, à savoir un estimateur comportant peu de calculs et adapté à un échantillon non probabiliste. Dans la section qui suit, nous examinons des méthodes qui permettront de redresser les estimations régionales de manière à faire concorder leur somme avec l'estimation de la production totale de l'Etat calculée par le DAEU.

#### 4. PONDERATION DES ESTIMATIONS EN VUE D'UNE CONCORDANCE AVEC LE TOTAL ESTIMÉ POUR L'ÉTAT

Soit  $Y$  l'estimation de la production de blé au Kansas, telle qu'établie par le DAEU. En règle générale,  $\sum_{i=1}^I Y_{i+} \neq Y$ . Par conséquent, nous définissons une nouvelle estimation

$$\tilde{Y}_{i+} = c_i Y_{i+},$$

où les  $c_i$  sont des constantes telles que  $\sum_{i=1}^I \tilde{Y}_{i+} = \sum_{i=1}^I c_i Y_{i+} = Y$ . La question qu'il faut se poser maintenant est de savoir comment on choisit les valeurs  $c_i$ . Selon les méthodes en usage actuellement, on pose  $c_i = c$  (au niveau du district) et on pondère par conséquent toutes les estimations par le même facteur. Une autre façon de procéder est de choisir  $c_i$  de manière à minimiser la somme des carrés des écarts ou des carrés des écarts relatifs entre  $\tilde{Y}_{i+}$  et  $Y_{i+}$ . Nous définissons ci-dessous les valeurs de  $c_i$  et de  $\tilde{Y}_{i+}$  pour les trois scénarios possibles.

##### 1) Poser $c_i = c$

Si on définit  $c_i$  comme une constante, il est facile de montrer que

$$c_i = c = Y / \sum_{i=1}^I Y_{i+}$$

et

$$\tilde{Y}_{i+} = Y \left( Y_{i+} / \sum_{i=1}^I Y_{i+} \right).$$

##### 2) Choisir $c_i$ de manière à minimiser la somme des carrés des écarts entre $\tilde{Y}_{i+}$ et $Y_{i+}$

Si nous voulons choisir les  $c_i$  de manière à minimiser la somme des carrés des écarts entre  $\tilde{Y}_{i+}$  et  $Y_{i+}$ , pourvu que  $\sum_{i=1}^I c_i Y_{i+} = Y$ , nous devons minimiser  $\sum_{i=1}^I (\tilde{Y}_{i+} - Y_{i+})^2 = \sum_{i=1}^I (c_i Y_{i+} - Y_{i+})^2$  par rapport aux valeurs des  $c_i$  au moyen d'un multiplicateur de Lagrange afin d'imposer la condition voulue. De cette manière, nous observons que

$$c_i = 1 + \left[ Y - \sum_{i=1}^I Y_{i+} \right] \left( Y_{i+}^2 / \sum_{i=1}^I Y_{i+}^2 \right) / \left( 1 / Y_{i+} \right)$$

et

$$\tilde{Y}_{i+} = Y_{i+} + \left[ Y - \sum_{i=1}^I Y_{i+} \right] \left( Y_{i+} / \sum_{i=1}^I Y_{i+} \right) \left( 1 / Y_{i+} \right).$$

On peut alors estimer la production du  $i$ -ième comté par la formule:

$$Y_{i+} = \sum_{N_i}^{N_i} Y_{ij} = \sum_{N_i}^{N_i} f(X_{ij} | \hat{\beta}),$$

où le “+” en indice signifie que la sommation est étendue à tous les  $j$ . Si nous voulions définir une forme générale pour  $f(X_{ij} | \beta)$ , il nous faudrait connaître la valeur de  $X_{ij}$  pour toutes les exploitations du  $i$ -ième comté. Or, nous savons pertinemment qu’il n’est pas possible d’obtenir des renseignements aussi détaillés. Cependant, si  $f(X_{ij} | \beta)$  est une fonction linéaire, il nous suffit de connaître la valeur totale des variables explicatives pour chaque comté. Tel est le cas puisque pour une équation de régression linéaire,

$$Y_{i+} = \sum_{N_i}^{N_i} Y_{ij} = \sum_{N_i}^{N_i} [\hat{\beta}_0 + \hat{\beta}_1 X_{ij1} + \hat{\beta}_2 X_{ij2} + \dots + \hat{\beta}_p X_{ijp}]$$

$$= \hat{\beta}_0 N_i + \hat{\beta}_1 X_{i+1} + \hat{\beta}_2 X_{i+2} + \dots + \hat{\beta}_p X_{i+p},$$

où  $X_{i+k}$  est la valeur totale de la  $k$ -ième variable explicative pour le  $i$ -ième comté.

Les valeurs  $Y_{i+}$  seront des estimations raisonnables si le modèle de régression décrit la relation entre les variables explicatives et la production totale des fermes de chaque comté de même que la production des fermes de la base de données. Toutefois, la somme de ces estimations ne correspondra pas nécessairement à l’estimation de la production totale de l’Etat calculée par le DABU. Dans la section 4, nous envisageons des méthodes qui permettront de résoudre ce problème.

Le modèle de régression proposé ci-dessus permet de calculer non seulement des estimations de la production agricole par comté, mais aussi des estimations de la variance. On peut le voir très facilement en exprimant les estimations de comté sous forme de matrices. Soit

$X$  = matrice  $n \times (p + 1)$  des données réelles, où les lignes sont formées des valeurs  $X_{ij}$  définies plus haut;

$Z$  = matrice (inconnue)  $N \times (p + 1)$  des variables explicatives pour toutes les exploitations agricoles d’un Etat;

$Y$  = vecteur (inconnu)  $N \times 1$  des estimations de la production de blé pour les  $N$  exploitations d’un Etat;

$B_i$  = vecteur colonne  $N \times 1$  composé des éléments  $b_{ij}$ ;

$$\text{où } b_{ij} = \begin{cases} 1 & \text{si la } j\text{-ième ferme se trouve dans le } i\text{-ième comté} \\ 0 & \text{dans le cas contraire;} \end{cases}$$

$$A = [B_1 B_2 B_3 \dots B_I]^{N \times I}.$$

La méthode d’estimation décrite plus haut ne donne pas  $Y$  mais un vecteur d’estimations régionales  $Y_{i+}$ , où “ $Y$ ” désigne la transposée d’une matrice. La variance des estimations par comté est donc

$$\text{Var}(Y_{i+}) = \text{Var}(A^T Y) = A^T \text{Var}(Y) A = A^T \text{Var}(Z \hat{\beta}) A = \sigma^2 A^T Z (X^T X)^{-1} Z^T A.$$



### 3. MODÈLE DE RÉGRESSION

Nous proposons d'élaborer un modèle de régression qui servira à produire des estimations par comté. Il existe un certain nombre de logiciels statistiques conçus pour les ordinateurs personnels et qui permettent l'ajustement de modèles de régression multiple. En outre, l'estimateur que nous proposons tient compte de ce que nous n'utilisons pas un échantillon aléatoire et produira des estimations régionales dont la somme concorde avec le total estimé pour l'État.

Voici comment nous allons procéder:

- 1) Décrire par un modèle de régression multiple la relation entre la production agricole et certaines variables explicatives en se servant de l'échantillon non probabiliste de fermes.
- 2) Supposer que cette relation vaut pour toutes les fermes sur le territoire de l'État et estimer la production agricole totale dans chaque comté.
- 3) Redresser les estimations par comté de la production agricole de manière que leur somme concorde avec l'estimation calculée par le DABU pour l'État.

Pour décrire le modèle de régression, nous allons utiliser la notation suivante. Pour  $i = 1, 2, \dots, I(I = 105 - \text{nombre de comtés au Kansas})$  et  $j = 1, 2, \dots, n_j$ , posons

$n_i$  = nombre de fermes du  $i$ -ième comté dans l'échantillon;

$$n = \sum_{i=1}^I n_i = \text{taille de l'échantillon global};$$

$N_i$  = nombre total de fermes dans la  $i$ -ième comté dans la population;

$$N = \sum_{i=1}^I N_i = \text{nombre total de fermes dans la population};$$

$Y_{ij}$  = production de blé de la  $j$ -ième ferme du  $i$ -ième comté (en boisseaux);

$$X_{ij} = (1 \ X_{ij1} \ X_{ij2} \ \dots \ X_{ijp}) = \text{vecteur de } p \text{ variables explicatives pour la } j\text{-ième ferme du } i\text{-ième comté}.$$

Comme nous le verrons plus loin, il importe de choisir des variables explicatives pour lesquelles il existe des totaux de comté ou des estimations très précises de ces totaux. De plus, les variables explicatives doivent renfermer des éléments d'information qui ont trait à la probabilité d'échantillonnage d'une exploitation agricole, comme une mesure de la taille de l'exploitation. De cette manière, nous pouvons utiliser le modèle de régression en tenant compte de ce qu'il ne s'agit pas d'un échantillon aléatoire.

Nous considérons des modèles de régression de la forme

$$Y_{ij} = f(X_{ij} | \beta) + \epsilon_{ij},$$

où  $\beta = (\beta_0 \ \beta_1 \ \beta_2 \ \dots \ \beta_p)$  est un vecteur de paramètres et  $\epsilon_{ij}$  est un terme d'erreur aléatoire de variance  $\sigma^2$ . Nous allons représenter les valeurs ajustées, obtenues à l'aide de données de la base sur les menus grains, par l'expression suivante:

$$\hat{Y}_{ij} = f(X_{ij} | \hat{\beta}).$$

2. DONNÉES DU KANSAS

Aux fins de l'enregistrement des données sur la production agricole, chaque Etat est divisé en neuf ou dix districts. Le Kansas est découpé en neuf districts et en 105 comtés. Le tableau ci-dessous donne la situation géographique de chaque district ainsi que le nombre de comtés correspondant :

District numéro	Situation géographique	Nombre de comtés
1	Nord-Ouest	8
2	Centre ouest	9
3	Sud-Ouest	14
4	Centre nord	11
5	Centre	11
6	Centre sud	13
7	Nord-Est	11
8	Centre est	14
9	Sud-Est	14

Pour notre étude, nous disposons de deux bases de données qui servent à la production d'estimations pour les comtés du Kansas: la base de données sur les emblavures (Planted Acres Data Base) et la base de données sur les menus grains (Small Grain Data Base). Nous nous sommes servis principalement de données de 1987 mais nous avons validé nos résultats à l'aide des données de 1988. La base de données sur les emblavures pour 1987 contient des renseignements touchant 37,094 fermes réparties sur le territoire du Kansas. (Selon la définition établie par le DAEU, une ferme est une entité qui vend pour \$1,000 ou plus de produits agricoles par année.) De ces 37,094 fermes, 22,300 comptaient des terres ensemençées en blé; toutes ces fermes sans exception ont été utilisées dans l'étude de simulation décrite dans la section 5. La base de données sur les menus grains pour 1987 contient des renseignements touchant 5,802 fermes où l'on cultive des menus grains. De ces 5,802 fermes, 1,707 produisent du blé et elles ont toutes servi pour notre étude.

Les enregistrements de la base de données sur les emblavures sont un mélange de données agricoles du Kansas qui ont été recueillies auprès de diverses sources à divers moments. Primo, on dresse une liste d'exploitations agricoles (nom et adresse) à l'aide des données recueillies par les conseillers agricoles. Ces données peuvent être révisées à la lumière des résultats des enquêtes agricoles trimestrielles (Quarterly Agricultural Surveys) et des rapports agricoles mensuels (Monthly Farm Reports). Pour les besoins des enquêtes agricoles trimestrielles, on utilise des échantillons systématiques stratifiés d'environ 2,600 exploitations. Le taux de réponse se situe autour de 80%. Quant au rapport agricole mensuel, il est rempli par environ 3,000 agriculteurs qui ont accepté de prêter leur concours. Le même agriculteur peut produire un rapport mensuel pendant de nombreuses années. Les données les plus récentes pour chaque poste se trouvent dans la base sur les emblavures et l'enregistrement relatif à une exploitation quelconque pour une année quelconque peut contenir des données de diverses sources.

La base de données de 1987 sur les menus grains contient des renseignements sur la superficie ensemençée, la superficie moissonnée et la production de boisseaux en ce qui a trait aux exploitations qui participent aux enquêtes agricoles trimestrielles et à l'enquête du Kansas sur les menus grains (Kansas Small Grain Survey). Pour ce qui est de l'édition 1987 de cette enquête, environ 6,000 questionnaires ont été postés à un échantillon aléatoire d'exploitations agricoles; à peu près la moitié de ces questionnaires ont été remplis puis retournés.

Outre le risque d'un biais dû à la non-réponse pour les données de la base sur les menus grains, il y a la probabilité réelle d'un biais de réponse. En effet, dans leur déclaration, les agriculteurs tendent à sous-estimer la production réelle. À cause de tous ces facteurs – échantillon non conforme, biais dû à la non-réponse, biais de réponse, nous avons jugé pertinent d'élaborer la méthode d'estimation décrite dans les sections qui suivent.

fonction de ce qu'il sait déjà des fermes contenues dans l'échantillon, des conditions atmosphériques et d'autres facteurs, puis il évalue l'effet de ces corrections sur l'estimation de la production totale pour l'Etat. Il peut répéter ce processus un certain nombre de fois, jusqu'à ce qu'il juge raisonnables les estimations pour chaque comté et jusqu'à ce que l'estimation de la production totale de l'Etat concorde avec celle établie par le DABU. (L'estimation calculée par le DABU repose sur un grand échantillon probabiliste et pour cette raison, elle est jugée plus précise que l'estimation globale obtenue en faisant la somme des estimations par comté. C'est pourquoi les Etats font généralement concorder la somme des estimations par comté avec l'estimation de la production totale établie par le DABU.)

De façon générale, il n'existe pas de documentation sur les méthodes d'estimation appliquées dans les Etats. Par conséquent, on ne peut vérifier la pertinence des hypothèses et des méthodes qu'utilise l'expert et il est pratiquement impossible d'analyser ces méthodes ou de refaire les calculs. En outre, on ne peut calculer des estimations de la variance ni appliquer la même méthode dans plus d'un Etat. L'élaboration de nouvelles méthodes d'estimation pour les comtés doit permettre de résoudre ces difficultés.

Les données qui ont servi à notre étude ont été recueillies au Kansas en 1987, soit avant que le DABU mette en application ses nouvelles techniques d'échantillonnage pour l'estimation par comté. Nous avons utilisé des données de 1987 car c'est l'année où s'est fait le Recensement de l'agriculture des Etats-Unis et il se pourrait donc que nous recourions à des données du recensement pour les besoins de l'estimation. Si nous avons choisi d'utiliser les données du Kansas, c'est que cet Etat a un programme de collecte des données régionales parmi les plus complets aux Etats-Unis. Néanmoins, les données qui servent à l'estimation par comté au Kansas, comme dans la plupart des autres Etats, ne proviennent pas d'un échantillon aléatoire d'exploitations agricoles. Par conséquent, il ne faudra pas utiliser un échantillon aléatoire de fermes à blé pour la méthode d'estimation que nous proposons. Du reste, cette méthode pourra être utile dans le cadre du nouveau programme d'estimation par comté puisque celui-ci n'oblige pas les Etats à utiliser des échantillons probabilistes.

L'estimation pour petites régions a fait l'objet de nombreuses études ces dernières années (voir, par exemple, Platak *et al.*, 1987). Selon les méthodes normales d'estimation pour petites régions, les probabilités d'échantillonnage doivent être connues puisqu'on se sert de l'inverse de la probabilité de sélection pour pondérer les observations dans les formules d'estimateurs courants comme l'estimateur synthétique et l'estimateur direct. (Voir, par exemple, Section 2 de Sarnadal et Hidiroglou (1989) pour une analyse des estimateurs pour petites régions.)

Les méthodes que nous allons examiner ici doivent différer des méthodes habituelles d'estimation pour petites régions parce que primo, l'échantillon de fermes utilisé n'est pas un échantillon aléatoire et secundo, la somme des estimations par comté doit concorder avec l'estimation de la production totale de l'Etat calculée par le DABU. De plus, comme le département de l'agriculture de la plupart des Etats ne dispose pas actuellement d'un gros ordinateur, les calculs pertinents doivent être suffisamment simples pour pouvoir être exécutés sur un ordinateur personnel. Par conséquent, nous éviterons, au départ, d'utiliser des estimateurs comportant de nombreux calculs comme ceux que décrivent Fay et Herriot (1979). C'est pourquoi nous examinerons un estimateur qui comporte peu de calculs et qui repose sur un modèle de régression.

Dans la section 2, nous décrivons les bases de données du Kansas qui ont servi à cette étude. La section suivante expose la méthode de régression utilisée pour estimer la production de blé tandis que la section 4 décrit plusieurs méthodes permettant de pondérer les estimations par régression de manière qu'elles concordent avec l'estimation de la production totale de l'Etat calculée par le DABU. Dans la section 5, nous comparons les estimations par régression aux estimations publiées (par comté) et à des estimations calculées au moyen de l'estimateur synthétique et de l'estimateur direct. La section 6 contient les résultats d'une étude de simulation qui visait à comparer les divers estimateurs précités. Enfin, la section 7 renferme les conclusions et propose des sujets de recherche.



## Estimation de la production de blé par comité

ELIZABETH A. STASNY, PREM K. GOEL  
et DEBORAH J. RUMSEY<sup>1</sup>

### RÉSUMÉ

Bien que les enquêtes agricoles du Département de l'agriculture des E.-U. (DAEU) servent à estimer la production végétale à l'échelle du pays et pour chaque État, les estimations par comité sont plus utiles aux décideurs locaux. Ce genre d'estimations sont aussi recherchées par les entreprises qui se spécialisent dans la vente d'engrais, de pesticides, d'assurance-récolte et de matériel agricole. Les États font souvent leurs propres enquêtes dans le but de produire des estimations de la production agricole par comité. En règle générale, ces enquêtes ne reposent pas sur des méthodes d'échantillonnage probabiliste. En outre, la somme des estimations pour chaque comité d'un État doit concorder avec l'estimation calculée par le DAEU pour l'ensemble de l'État. Les méthodes d'estimation classiques pour petites régions ne sont donc pas directement applicables dans les circonstances. Dans cet article, nous allons recourir à des modèles de régression pour estimer la production de blé par comité au Kansas. Nous allons décrire une étude de simulation par laquelle nous comparons les estimations obtenues par régression à celles calculées à l'aide de deux estimateurs classiques pour petites régions: l'estimateur synthétique et l'estimateur direct. Nous allons aussi comparer plusieurs méthodes par lesquelles nous pouvons pondérer les estimations initiales de manière qu'elles concorderont avec l'estimation de la production totale de l'État calculée par le DAEU.

MOTS CLÉS: Échantillon non-probabiliste; régression; simulation; estimation pour petits domaines.

### 1. INTRODUCTION

Les estimations de la production agricole par comité sont de plus en plus recherchées par les organismes publics, qui s'en servent dans l'élaboration des décisions économiques au niveau local, et par les entreprises qui vendent des engrais, des pesticides, de l'assurance-récolte et du matériel agricole. Le Département de l'agriculture des États-Unis (DAEU) est en train de mettre sur pied un programme visant à uniformiser et à améliorer les méthodes d'estimation de la production agricole par comité (Bass et coll., 1989). Jusqu'à maintenant, chaque État avait sa façon propre d'établir des estimations par comité, c'est ce qui explique le peu d'uniformité des méthodes de collecte de données et d'estimation des divers États. Par son programme, le DAEU cherche à mettre à la disposition des États un ensemble de méthodes d'échantillonnage et d'estimation qui favorise une uniformisation de la qualité des estimations dans tout le pays.

Le nouveau programme du DAEU touche tous les aspects de la production d'estimations régionales (ou par comité), depuis la construction de bases de sondage jusqu'à l'estimation proprement dite. En ce qui concerne la présente étude, nous nous penchons uniquement sur l'estimation de la production de boisseaux de blé. Nous espérons toutefois que les méthodes proposées pourront servir à d'autres volets, comme l'estimation de la superficie ensemencée ou de la production d'autres cultures que le blé.

Malgré la diversité des méthodes d'estimation qu'utilisent les États, on note des points communs. Ainsi, dans chaque État, on calcule des estimations initiales à partir des données recueillies dans chaque comité. Ensuite, un expert examine ces estimations, les corrige en

<sup>1</sup> Elizabeth A. Stasny, Prem K. Goel et Deborah J. Rumsey, Department of Statistics, The Ohio State University, 1958 Neil Avenue, Columbus, Ohio 43210, USA.



Les auteurs tiennent à exprimer leur reconnaissance à Lyne Guertin, qui a programmé l'étude de simulation. Ils remercient également les arbitres et J.N.K. Rao pour les commentaires utiles et constructifs qu'ils leur ont faits.

## BIBLIOGRAPHIE

- BANKIER, M.D. (1988). Power allocations: Determining sample sizes for subnational areas. *The American Statistician*, 42, 174-177.
- BREWER, K.R.W., EARLY, L.J., et HANIF, M. (1984). Poisson, modified Poisson and collocated sampling. *Journal of Statistical Planning and Inference*, 10, 15-30.
- BREWER, K.R.W., EARLY, L.J., et JOYCE S.F. (1974). Selecting several samples from a single population. *Australian Journal of Statistics*, 14, 232-239.
- COCHRAN, W.G. (1977). *Sampling Techniques*, (3ième édition). New York: John Wiley.
- DALENIUS, T., et HODGES, J.L., Jr. (1959). Minimum variance stratification. *Journal of the American Statistical Association*, 54, 88-101.
- EARLY, L.J., et BREWER, K.R.W. (1971). Some estimators for arbitrary probability sampling. Master's thesis.
- HAIJK, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics*, 35, 1491-1523.
- HANSEN, M.H., HURWITZ, W.N., et MADOW, W.G. (1953). *Sample Survey Methods and Theory*. New York: John Wiley and Sons.
- HIDIROGLOU, M.A. (1986). The construction of a self-representing stratum of large units in survey design. *The American Statistician*, 40, 27-31.
- KISH, L., et SCOTT, A. (1971). Retaining units after changing strata and probabilities. *Journal of the American Statistical Association*, 66, 461-470.
- MICKEY, M.R. (1959). Some finite population unbiased ratio and regression estimators. *Journal of the American Statistical Association*, 54, 594-612.
- RAO, J.N.K., et KUZIK, R.A. (1974). Sampling errors in ratio estimation. *Sankhyā. Series, C*, 36, 43-58.
- SUNTER, A.B. (1977). Response burden, sample rotation and classification renewal in economic surveys. *International Statistical Review*, 45, 209-222.

## REMERCIEMENTS



Tableau 3a  
Efficacité relative (EFF) en pourcentage  
Niveau global

Scénario	Valeurs dilatées	Quotient séparé	Quotient combiné	Mickey
1	100.0	108.0	107.9	107.3
2	100.0	100.2	99.8	100.1
3	100.0	148.3	160.3	143.5
4	100.0	74.3	92.3	84.3

Tableau 3b  
Efficacité relative moyenne (EFF) en pourcentage  
Niveau de la strate

Scénario	Valeurs dilatées	Quotient séparé	Quotient combiné	Mickey
1	100.0	109.6	108.6	108.6
2	100.0	100.9	99.5	100.6
3	100.0	183.3	180.2	174.2
4	100.0	80.0	99.4	83.7

i) estimateur par quotient séparé, ii) estimateur par quotient combiné, iii) estimateur de Mickey et iv) estimateur à valeurs dilatées. Enfin, pour ce qui a trait au scénario 4, qui représente la pire situation possible (déséquilibre des effectifs et corrélation irrégulière), le meilleur estimateur, tant au niveau global qu'au niveau de la strate, est l'estimateur à valeurs dilatées. L'estimateur de Mickey a permis d'observer des poids négatifs dans 2% des cas. En conclusion, étant donné les quatre scénarios ci-dessus, nous pouvons dire que l'estimateur par quotient combiné est un choix convenable pour les enquêtes infra-annuelles qui reposent sur un plan de sondage avec groupes de renouvellement. L'estimateur à valeurs dilatées mérite peut-être aussi d'être considéré en raison de sa simplicité. Cependant, il faut se rappeler que ses propriétés conditionnelles laissent à désirer lorsque les effectifs des groupes de renouvellement ne sont pas équilibrés.

5. CONCLUSION

Dans cet article, nous avons décrit un plan de sondage qui peut répondre aux exigences d'une enquête infra-annuelle auprès d'entreprises. Ces exigences recouvrent l'échantillonnage initial de même que le renouvellement et la mise à jour de l'échantillon. Étant donné ce plan de sondage, nous avons examiné un certain nombre de méthodes d'estimation que nous avons évaluées au moyen d'une étude par simulation. Ces méthodes sont équivalentes lorsque les effectifs des groupes de renouvellement dans chaque strate sont équilibrés. S'il y a déséquilibre entre les effectifs, il est recommandé d'opter pour l'estimateur par quotient combiné, qui utilise la taille des groupes de renouvellement comme information supplémentaire.

Tableau 1  
Biens relatifs en valeur absolue (ARB) en pourcentage

Scénario	Niveau global		Niveau de la strate	
	Quotient séparé	Quotient combiné	Quotient séparé	Quotient combiné
1	1.27	0.07	1.31	0.11
2	0.02	0.01	0.24	0.05
3	2.88	0.14	3.19	0.29
4	5.51	0.22	5.72	0.30

Tableau 2  
Biens type absolu (ASB) en pourcentage

Scénario	Niveau global		Niveau de la strate	
	Quotient séparé	Quotient combiné	Quotient séparé	Quotient combiné
1	13.41	0.76	3.37	0.24
2	0.44	0.26	0.58	0.25
3	45.64	2.13	12.11	0.69
4	43.29	1.96	9.88	0.71

de la strate. Comme l'indique le tableau 1, le biais relatif le plus élevé est associé à l'estimateur par quotient séparé tandis que le biais relatif le plus faible est associé à l'estimateur par quotient combiné. Dans le cas de ces estimateurs, le biais relatif en valeur absolue augmente à mesure que s'accroît le coefficient de variation des effets des groupes de renouvellement et que diminue la corrélation entre ces effets et la variable étudiée.

En ce qui regarde le biais type absolu (ASB), le tableau 2 nous permet de faire les observations suivantes. L'estimateur par quotient séparé est inacceptable pour la plupart des scénarios. Son rendement diminue à mesure que le coefficient de variation des effets des groupes de renouvellement augmente et que la corrélation entre ces effets et la variable étudiée diminue. En revanche, l'estimateur par quotient combiné est acceptable dans tous les cas.

L'efficacité relative (EFF) des estimateurs est exposée dans les tableaux 3a et 3b. Dans le cas du scénario 1, où il y a un bon équilibre des effets des groupes de renouvellement et une forte corrélation entre ces effets et la variable étudiée, tous les estimateurs sont équivalents, au niveau global comme au niveau de la strate. Pour ce qui est du scénario 2, où l'équilibre des effets est bon et la corrélation irrégulière, les estimateurs sont aussi équivalents aux deux niveaux. Quant au scénario 3, caractérisé par un déséquilibre entre les effets des groupes de renouvellement et une forte corrélation entre ces effets et la variable étudiée, les estimateurs se classent dans l'ordre suivant au niveau global (de l'efficacité relative la plus élevée à la moins élevée): i) estimateur par quotient combiné, ii) estimateur par quotient séparé, iii) estimateur de Mickey et iv) estimateur à valeurs diluées. Au niveau de la strate, ce classement devient:

Les quatre scénarios envisagés représentent des situations qui peuvent survenir au fil de l'enquête. Le scénario 1 correspond à la situation que l'on observe au moment de l'échantillonnage initial: l'équilibre des effectifs des groupes de renouvellement est bon et la corrélation entre ces effectifs et la variable étudiée est forte. Le scénario 2 traduit la diminution de la corrélation que l'on observe au fil du temps à cause des unités mortes qui s'accroissent au sein de la population. Comme, dans ce cas, les unités mortes n'ont pas été supprimées de la population, l'équilibre des effectifs des groupes de renouvellement est bon mais la corrélation entre ces effectifs et la variable étudiée n'est pas aussi forte qu'au début. Le scénario 3 implique que la suppression des unités disparues créera vraisemblablement un déséquilibre entre les effectifs des groupes de renouvellement mais accroîtra la corrélation entre ces effectifs et la variable étudiée. Enfin, le scénario 4 représente la pire situation possible, soit une corrélation faible entre l'effectif des groupes de renouvellement et la variable étudiée et un mauvais équilibre des effectifs des groupes.

Nous avons élaboré le scénario 1 en variant les effectifs des groupes de renouvellement sans changer les valeurs  $y_{hi}$  (RBE). Nous avons donc fixé la taille des 16 groupes de renouvellement en les classant, quatre par quatre, par ordre croissant de la valeur de  $y_{hi}$ . Ainsi, les effectifs des groupes de renouvellement 1 à 4, 5 à 8, 9 à 12 et 13 à 16 étaient, respectivement,  $0,22 M_h/4$ ,  $0,24 M_h/4$ ,  $0,26 M_h/4$  et  $0,28 M_h/4$ . La corrélation moyenne entre le RBE et la taille des groupes de renouvellement était de  $0,86$ , avec un intervalle de variation de  $0,69$  à  $0,96$  au niveau de la strate. Le coefficient de variation moyen des effectifs des groupes de renouvellement était de  $9,2\%$ .

En ce qui concerne le scénario 2, nous avons permuté aléatoirement les unités de la population puis nous les avons réparties systématiquement entre 16 groupes de renouvellement selon la méthode décrite dans la section 2.2. Ensuite, nous avons imputé aléatoirement une valeur  $y$  nulle à environ  $20\%$  de ces unités pour tenir compte de la proportion des unités disparues. La corrélation globale entre le RBE et la taille des groupes de renouvellement était de  $0,11$ , avec un intervalle de variation de  $-0,23$  à  $0,74$  au niveau de la strate. Le coefficient de variation moyen des effectifs des groupes de renouvellement était de  $4,1\%$ .

Pour ce qui a trait au scénario 3, nous avons procédé de la même manière que pour le scénario 1 sauf que les effectifs des groupes de renouvellement étaient différents. Ainsi, les effectifs des groupes 1 à 4, 5 à 8, 9 à 12 et 13 à 16 étaient, respectivement,  $0,05 M_h/4$ ,  $0,20 M_h/4$ ,  $0,30 M_h/4$  et  $0,45 M_h/4$ . La corrélation globale entre le RBE et la taille des groupes de renouvellement était de  $0,87$ , avec un intervalle de variation de  $0,70$  à  $0,96$  au niveau de la strate. Le coefficient de variation moyen des effectifs des groupes de renouvellement était de  $60,2\%$ .

Quant au scénario 4, nous avons déterminé aléatoirement la taille des groupes de renouvellement en faisant abstraction de la variable RBE. Nous avons supposé que, pour chaque strate  $h$ ,  $a_h = \min\{M_{hi}; i = 1, \dots, N_h\}$  et  $b_h = \max\{M_{hi}; i = 1, \dots, N_h\}$ . Pour chaque strate  $h$ , nous avons défini la taille  $M_{hi}^*$  du groupe de renouvellement  $i$  par l'expression  $n_h e_{hi}$ , où  $e_{hi}$  est distribué uniformément dans l'intervalle  $(a_h, b_h)$  et  $n_h$  est un facteur d'échelle tel que  $M_h = \sum_{i=1}^{N_h} M_{hi}^*$ . La corrélation moyenne était nulle, avec un intervalle de variation de  $-0,49$  à  $0,56$  au niveau de la strate. Le coefficient de variation moyen des effectifs des groupes de renouvellement était de  $49,2\%$ .

#### 4.4 Analyse des résultats

En nous servant des quatre scénarios décrits ci-dessus, nous avons calculé par simulation le biais relatif en valeur absolue (ARB), le biais type absolu (ASB), l'efficacité (EFF) et la proportion de poids nuls ou négatifs pour chaque strate et pour l'ensemble de l'échantillon. Les résultats figurent dans les tableaux 1 à 3. Notons que tous ces résultats sont exprimés en pourcentage.

En ce qui a trait au biais relatif en valeur absolue (ARB), l'estimateur à valeurs diluées et l'estimateur de Mickey en sont tous deux exempts, que ce soit au niveau global ou au niveau



et le biais relatif global au moyen de la formule

$$ARB = \left| \sum_{h=1}^L \text{Biais}(Y_h) \right| / Y,$$

où

$$Y = \sum_{h=1}^L Y_h.$$

Le deuxième critère était le rapport entre le biais en valeur absolue et l'erreur type, ce rapport étant appelé "biais type absolu". Nous avons calculé le biais type absolu moyen de strate à l'aide de la formule

$$\underline{ASB} = \frac{1}{L} \sum_{h=1}^L \left| \text{Biais}(Y_h) \right| / \sqrt{\text{Var}(Y_h)}$$

et le biais type absolu global au moyen de la formule

$$ASB = | \text{Biais}(Y) | / \sqrt{\text{Var}(Y)}.$$

D'après Cochran (1977), le biais type absolu ne devrait pas excéder 10%. En effet, comme la précision d'un estimateur est mesurée habituellement par la variance et non par l'EQM (MSE), un biais trop élevé par rapport à l'écart type aurait pour effet de créer une fausse idée de la précision de l'estimateur utilisé.

Le troisième critère était l'efficacité, définie comme le rapport entre la racine carrée de l'EQM de l'estimateur à l'étude,  $\text{RMSE}(Y_{EST})$ , et la racine carrée de l'EQM de l'estimateur à valeurs diluées,  $\text{RMSE}(Y_{EXP})$ . Nous avons calculé l'efficacité relative moyenne de strate à l'aide de la formule

$$\overline{EFF} = \frac{1}{L} \sum_{h=1}^L \left\{ \text{RMSE}(Y_{EXP}^h) / \text{RMSE}(Y_{EST}^h) \right\},$$

et l'efficacité relative globale au moyen de la formule

$$EFF = \text{RMSE}(Y_{EXP}) / \text{RMSE}(Y_{EST}).$$

Enfin, le quatrième critère était la proportion de poids négatifs.

### 4.3 Description des scénarios

Nous avons envisagé quatre scénarios relativement à la configuration de la population de groupes de renouvellement formée pour les besoins du plan d'échantillonnage décrit dans la section 2.2. Les quatre scénarios correspondaient à autant de possibilités concernant l'équilibre des effectifs des groupes de renouvellement (bon, mauvais) et la corrélation entre l'effectif des groupes de renouvellement,  $M_{hi}$ , et la variable étudiée,  $y_{hi}$  (forte, irrégulière). Pour ce qui a trait à l'équilibre des effectifs des groupes de renouvellement, le qualificatif "bon" signifie que ces effectifs sont comparables tandis que "mauvais" signifie qu'ils diffèrent sensiblement les uns des autres. En ce qui concerne la corrélation, le qualificatif "forte" signifie que la corrélation entre la variable étudiée et la taille des groupes de renouvellement est très élevée dans toutes les strates alors que "irrégulière" signifie qu'elle varie de faible à élevée selon les strates.

$$E(Y_h) = \frac{1}{K} \sum_{k=1}^K Y_{h(k)},$$

où  $K$  est le nombre total d'échantillons prélevés. Notons que pour les estimateurs (3.1), (3.2) et (3.4),  $E(Y_h)$  était en fait l'espérance puisque tous les échantillons possibles avaient été prélevés. Dans le cas de l'estimateur (3.3) - estimateur par quotient combiné -  $E(Y_h)$  était plutôt une estimation sans biais de l'espérance. Le biais de strate était défini

$$\text{Biais}(Y_h) \doteq E(Y_h) - Y_h.$$

On calculait le biais global, Biais  $(Y)$ , en faisant la somme des biais de strates. Pour les estimateurs (3.1), (3.2) et (3.4), nous avons

$$\text{Var}(Y_h) \doteq \frac{1}{K} \sum_{k=1}^K (Y_{h(k)} - E(Y_h))^2$$

et

$$\text{Var}(Y) = \sum_L^h \text{Var}(Y_h).$$

Pour l'estimateur par quotient combiné (3.3), nous avons

$$\text{Var}(Y_h) \doteq \frac{K-1}{1} \sum_K^k (Y_{h(k)} - E(Y_h))^2$$

et

$$\text{Var}(Y) = \frac{K-1}{1} \sum_K^k (Y_{(k)} - E(Y))^2,$$

où  $Y_{(k)} = \sum_{h=1}^L Y_{h(k)}$  et  $E(Y) = \sum_{h=1}^L E(Y_h)$ .

Enfin, pour chaque estimateur, l'EQM de strate,  $\text{MSE}(Y_h)$ , était définie

$$\text{MSE}(Y_h) = \text{Var}(Y_h) + (\text{Biais}(Y_h))^2$$

tandis que l'EQM globale,  $\text{MSE}(Y)$ , était définie

$$\text{MSE}(Y) = \text{Var}(Y) + (\text{Biais}(Y))^2.$$

Pour comparer les estimateurs proposés entre eux, nous avons servis de quatre critères, le premier étant le biais relatif en valeur absolue. Nous avons calculé le biais relatif moyen de strate (en valeur absolue) à l'aide de la formule

$$\overline{\text{ARB}} = \frac{1}{L} \sum_{h=1}^L \left| \frac{\text{Biais}(Y_h)}{Y_h} \right|$$

Un estimateur jackknife de la variance de  $X_{MI,h}(d)$  est défini

$$v_j(X_{MI,h}(d)) = (1 - f_h) \frac{n_h}{\sum_{h=1}^J (z_h^{(j)}(d) - z_h(d))^2},$$

$$\text{ou } z_h^{(j)}(d) = Y_{MI,h}^{(j)}(d) \text{ et } z_h(d) = n_h^{-1} \sum_{j=1}^J z_h^{(j)}(d).$$

On peut montrer que tous les estimateurs sont équivalents et inconditionnellement sans biais lorsque les groupes de renouvellement ont tous la même taille  $M_h$  dans chaque strate  $h$ . Cependant, lorsque ce n'est pas le cas, tous les estimateurs, sauf l'estimateur à valeurs diluées et l'estimateur de Mickey, sont inconditionnellement biaisés. Dans la section suivante, nous présentons une étude par simulation qui vise à mesurer l'importance du biais inconditionnel et l'efficacité de ces estimateurs.

### 4. ETUDE PAR SIMULATION

Cette simulation avait pour but de déterminer lequel des quatre estimateurs de l'agrégat  $Y(d)$  et du total de strate  $Y_h(d)$  pouvait "convenir" le mieux au plan de sondage décrit dans la section 2. Par souci de simplicité, nous avons limité les simulations à une seule variable ( $y$ ), soit le revenu brut de l'entreprise (RBE). De plus, pour les besoins de cette simulation, nous avons fait coïncider les domaines avec les strates. Par conséquent, nous omettrons le symbole "d" qui sert à désigner les domaines.

#### 4.1 Description de l'étude

Nous avons défini l'univers de cette étude comme l'ensemble des unités de petite taille qui, en mai 1989 (période de référence), appartenaient au secteur du commerce de gros de la province de Québec. La taille de chaque unité a été établie en fonction du RBE que l'on avait déduit des comptes de retenues sur la paye à l'aide d'un modèle fondé sur un rapport. Les unités dont le RBE était inférieur à un seuil donné étaient retenues pour faire partie de l'univers; nous avons ainsi constitué une population de 10,953 unités. Nous avons ensuite stratifié cette population selon le niveau à 3 chiffres de la Classification type des industries. Cette opération a produit 30 strates contenant au moins 18 unités chacune. Ensuite, nous avons formé 16 groupes de renouvellement à l'intérieur de chaque strate en répartissant aléatoirement les unités de la façon décrite dans la section 2.2.2.

Pour chaque strate  $h$ , nous avons prélevé 4 groupes de renouvellement parmi les 16 au moyen d'un échantillonnage aléatoire simple sans remise. Cela signifie que l'on pouvait tirer jusqu'à 1,820 échantillons différents par strate. En ce qui concerne l'estimation par quotient séparé, cela donne 54,600 estimations possibles (30 strates  $\times$  1,820 échantillons par strate). Quant à l'estimateur par quotient combiné, on pourra avoir jusqu'à (1,820)<sup>30</sup> estimations différentes. Dans le cas de l'estimateur à valeurs diluées, de l'estimateur par quotient séparé et de l'estimateur de Mickey, nous avons tiré les 54,600 échantillons possibles. Pour ce qui est de l'estimateur par quotient combiné, nous avons tiré au hasard 100,000 échantillons parmi les (1,820)<sup>30</sup> échantillons possibles.

#### 4.2 Critères d'évaluation

Les critères d'évaluation avaient trait au biais et à l'EQM. Nous les décrivons ci-dessous. Pour chaque échantillon  $k$ , prélevé, on a produit une estimation  $Y_h^{(k)}$  pour chaque strate  $h$  et pour chacun des quatre estimateurs. L'espérance de strate  $E(Y_h^{(k)})$  de cette estimation était calculée au moyen de l'expression



Cet estimateur a toutefois un inconvénient majeur: il est exposé au biais propre à l'estimation par quotient. Par conséquent, si le biais tend à être positif (ou négatif) dans la majorité des strates, son effet cumulatif pourra être très appréciable lorsque la somme sera étendue à toutes les strates.

C. Estimateur par quotient combiné

On peut réduire sensiblement l'effet cumulatif du biais d'agrégation en utilisant une version combinée de l'estimateur par quotient. L'estimateur par quotient combiné est défini par l'expression

$$Y^{CR}(d) = M \frac{\sum_{h=1}^L N_h y_h(d)}{\sum_{h=1}^L N_h m_h}, \tag{3.3}$$

où  $M = \sum_{h=1}^L M_h$ .

D. Estimateur par quotient sans biais

On peut supprimer le biais dû à l'estimation par quotient en utilisant une version de l'estimateur par quotient proposée par Mickey (1959). L'estimateur de Mickey est défini

$$Y_{MI}(d) = \sum_L^h \left( f_h(d) M_h + (N_h - n_h + 1) \left[ \sum_{n_h}^{y_{hi}(d)} y_{hi}(d) - m_h f_h(d) \right] \right), \tag{3.4}$$

où

$$f_h(d) = \frac{1}{n_h} \sum_{j=1}^J f_{hj}^h(d); f_{hj}^h(d) = \frac{\sum_{i \neq (j)}^{M_{hi}} y_{hi}(d)}{\sum_{i \neq (j)}^{M_{hi}} M_{hi}}; m_h = \sum_{n_h}^{M_{hi}} M_{hi}.$$

La faiblesse de cet estimateur est qu'il peut renfermer des poids inférieurs à un, y compris des poids négatifs.

Nous allons estimer la variance de l'estimateur par quotient séparé et celle de l'estimateur par quotient combiné à l'aide de la méthode de linéarisation de Taylor. Quant à la variance de l'estimateur de Mickey, nous allons nous servir de la méthode du jackknife. Pour cela, nous enlevons un groupe de renouvellement de l'échantillon puis calculons l'estimateur de Mickey pour les  $(n_h - 1)$  autres groupes; nous répétons ainsi l'opération pour chaque groupe. Désignons chaque estimateur soumis à la méthode du jackknife par  $Y_{MI,h}^{(j)}(d)$  pour  $j = 1, 2, \dots, n_h$ .

où

$$Y_{MI,h}^{(j)}(d) = \sum_{i \neq (j)}^{M_{hi}} w_{hi}^{(j)} y_{hi}(d)$$

avec

$$w_{hi}^{(j)} = [M_h - (m_h - M_{hj})] (N_h - n_h + 2) b_{hj}^{(j)} + (N_h - n_h + 2)$$

et

$$b_{hj}^{(j)} = (n_h - 1) \frac{\sum_{i \neq (j)}^{M_{hi}} (m_h - M_{hj})}{1}.$$

Nous allons maintenant examiner divers estimateurs du paramètre de population  $Y(d)$  et leurs variances respectives. Ces estimateurs s'écrivent sous la forme:

$$Y_h(d) = \sum_{h=1}^I w_{hi} y_{hi}(d) ,$$

où  $w_{hi}$  est le produit du poids de sondage par un facteur de redressement qui correspond à la méthode d'estimation utilisée. On obtient l'estimateur de  $Y(d)$  en faisant la somme des  $Y_h(d)$  par rapport aux strates, c'est-à-dire,

$$Y(d) = \sum_{h=1}^H Y_h(d) .$$

### 3.1 Estimateurs d'un total

#### A. Estimateur à valeurs diluées:

Comme la probabilité d'échantillonnage d'un groupe de renouvellement dans la strate  $h$  est  $n_h/N_h$ , le poids de sondage est  $w_{hi} = N_h/n_h$  pour  $i = 1, 2, \dots, n_h, h = 1, 2, \dots, L$ . L'estimateur à valeurs diluées est défini

$$Y^E(d) = \sum_{h=1}^L N_h \bar{y}_h(d) ,$$

où

$$\bar{y}_h(d) = n_h^{-1} \sum_{n_h}^I y_{hi}(d) . \tag{3.1}$$

Comme nous l'avons mentionné précédemment, cet estimateur est inconditionnellement sans biais mais peut renfermer parfois un biais conditionnel élevé. En outre, il peut ne pas être très efficace du fait qu'il n'exploite pas l'information supplémentaire disponible, par ex.: la taille des groupes de renouvellement. Et même, il peut devenir de moins en moins efficace à mesure que s'accroît la variation de la taille des groupes de renouvellement sous l'effet de la suppression des unités disparues.

#### B. Estimateur par quotient séparé

S'il y a une forte corrélation entre  $y_{hi}(d)$  et la taille des groupes de renouvellement  $M_{hi}$ , on peut réaliser des gains en efficacité au moyen de l'estimateur par quotient séparé, qui est défini

$$Y^{SR}(d) = \sum_{h=1}^H \left( \frac{M_h}{\bar{y}_h(d)} \right) \bar{y}_h(d) \tag{3.2}$$

où

$$\bar{m}_h = n_h^{-1} \sum_{n_h}^I M_{hi}$$

et

$$M_h = \sum_{N_h}^I M_{hi} .$$

3. PONDERATION ET ESTIMATION

L'estimateur le plus élémentaire que l'on puisse utiliser avec le plan d'échantillonnage fondé sur des groupes de renouvellement que nous avons décrit dans la section 2.2 est l'estimateur à valeurs diluées (ou estimateur pour domaines). Bien que cet estimateur soit inconditionnellement sans biais, il peut renfermer un biais conditionnel élevé lorsque les effectifs des groupes de renouvellement ne sont pas équilibrés, cela pouvant être attribuable à la suppression des unités disparues. C'est pourquoi nous avons envisagé d'autres estimateurs, qui tiennent compte de l'information supplémentaire que représente la taille des groupes de renouvellement. Ce sont l'estimateur par quotient séparé et l'estimateur par quotient combiné. Le premier présente toutefois un inconvénient: son biais peut prendre des proportions notables avec l'accumulation des strates. L'estimateur par quotient combiné, lui, aura un biais négligeable mais produira vraisemblablement des estimations de strate à variance élevée. Nous avons donc évalué l'efficacité d'un estimateur par quotient séparé sans biais élaboré par Mickey (1959). Or, pour un estimateur, la propriété d'être sans biais n'existe qu'au prix d'un accroissement de la variance. Notre premier objectif est de déterminer lequel des estimateurs ci-dessus convient le mieux au plan d'échantillonnage décrit dans la section 2.2. Le biais et l'erreur quadratique moyenne (EQM) serviront de critères d'évaluation à cette fin. Dans le but de simplifier les comparaisons, nous supposons que les données fournies pour chaque unité échantillonnée sont valides. Comme nous l'avons mentionné plus tôt, la strate  $h$  ( $h = 1, 2, \dots, L$ ) est définie en fonction de trois facteurs de classification: branche d'activité, région géographique et taille de l'entreprise. Il faut établir des estimations pour des domaines qui peuvent embrasser toutes les strates d'échantillonnage ou qui peuvent en être un sous-ensemble. Un exemple de tels domaines est l'agrégation de variables d'intérêt au niveau infra-provincial lorsque l'échantillonnage a été effectué à un niveau supérieur (au niveau provincial par exemple). Une caractéristique souhaitable est que la somme des estimations d'un ensemble de domaines disjoints soit toujours égale à l'estimation pour le domaine défini comme l'union de ces domaines. Pour qu'il y ait cohérence, on ne peut utiliser qu'une série de poids. Posons  $y$  comme la caractéristique étudiée et  $y_{hij}$  comme la valeur de cette caractéristique pour l'unité  $j$  du groupe de renouvellement (grappe)  $i$  de la strate  $h$ . Soit  $\delta_{hij}(d)$  une variable indicatrice qui prend la valeur 1 si l'unité  $hij$  appartient au domaine " $d$ " et la valeur 0 si ce n'est pas le cas. Alors, le paramètre qui nous intéresse est le total de population  $X(d)$ , qui est défini:

$$X(d) = \sum_{L=1}^h \sum_{N_h=1}^I \sum_{M_{hi}=1}^J y_{hij}(d),$$

$$\text{où } y_{hij}(d) = \delta_{hij}(d) y_{hij}.$$

Comme nous l'avons indiqué plus haut, nous avons un échantillon aléatoire simple de  $n_h$  groupes de renouvellement prélevés sans remise parmi les  $M_{hi}$  groupes de renouvellement de la strate  $h$ . Soit  $M_{hi}$  le nombre d'unités contenues dans le  $i$ -ième groupe de renouvellement de la strate  $h$ . Nous pouvons supposer, sans perte de généralité, que les groupes de renouvellement échantillonnés sont désignés par l'indice  $i = 1, 2, \dots, n_h$ . Soit  $y_{hi}(d)$  la réponse globale des unités appartenant au domaine " $d$ " dans le  $i$ -ième groupe de renouvellement échantillonné de la strate  $h$ , c.-à-d.

$$y_{hi}(d) = \sum_{M_{hi}=1}^{j=1} y_{hij}(d), i = 1, 2, \dots, n_h.$$



et

$$[(r - 1) \bmod P_k + 1, (r + P_k - 2) \bmod P_k + 1] \quad \text{si} \quad (r - 1) \bmod P_k \leq (P_k - P_k)$$

$$[1, P_k - P_k + (r - 1) \bmod P_k + 1, P_k] \quad \text{sinon.}$$

La première étape du rééchantillonnage consiste à transformer les divers intervalles d'échantillonnage, qui ont des bornes inférieures différentes, en des intervalles ayant une borne inférieure commune. Pour la strate  $k$ , l'intervalle d'échantillonnage correspondant est  $[1, P_k]$ . Posons  $b$  comme la limite inférieure de l'intervalle d'échantillonnage au temps  $t_2$ , définie  $(r - 1) \bmod P_k + 1$ . Toutes les unités qui ont pour numéro de renouvellement " $g$ " portent désormais le numéro  $(g - b + 1)$ , si  $b \leq g \leq P_k$ , ou le numéro  $P_k - (b - g - 1)$ , sinon.

La deuxième étape consiste à définir une nouvelle strate " $h$ " pour chaque unité de la population actuellement incluse dans la strate  $k$ . On peut donc représenter l'ensemble des unités de la nouvelle strate  $h$ ,  $U_h$ , comme l'union de  $K$  ensembles exclusifs et exhaustifs  $U_{hk}$ ,  $h = 1, 2, \dots, L$ . Chaque ensemble  $U_{hk}$  est formé des unités de population dont la nouvelle strate est  $h$  et la strate courante,  $k$ . Certains de ces ensembles peuvent être vides.

La troisième étape consiste à classer sur une échelle de 0 à 1 les unités d'échantillonnage de chaque ensemble  $U_{hk}$  en tenant compte de leur numéro de renouvellement. Supposons que l'ensemble  $U_{hk}$  contient  $M_{hk}$  unités dont les numéros de renouvellement varient de 1 à  $P_k$ . Classons ces unités de 1 à  $M_{hk}$  suivant le numéro de renouvellement attribué à chacune, les unités ayant les numéros les plus petits au bas de l'échelle et celles ayant les numéros les plus gros au haut de l'échelle. S'il existe des numéros égaux, on peut les brouiller aléatoirement en générant des nombres aléatoires uniformes. Ainsi, les unités de l'ensemble  $U_{hk}$  sont classées de 1 à  $M_{hk}$ . Dans un deuxième temps, on attribue à l'unité de rang " $i$ " dans l'ensemble  $U_{hk}$ , un nombre  $r_{hki} = (a_k + i - 1) / M_{hk}$ , où  $a_k$  est un nombre aléatoire généré uniformément dans l'intervalle  $[0, 1]$  pour chaque ensemble  $U_{hk}$  dans  $U_h$ . Le nombre  $r_{hki}$  représente le numéro de renouvellement de l'unité exprimé en fonction de l'intervalle  $[0, 1]$ . Supposons que la fraction de sondage rattachée à la nouvelle strate est  $f_h$  et que la fraction de sondage courante est  $f_k$ . Si  $f_h \geq f_k$ , alors toutes les unités de  $U_{hk}$  qui font partie de l'échantillon courant se retrouveront dans le nouvel échantillon, en plus des unités contenues dans l'intervalle fermé  $[0, f_h]$ . Si  $f_h < f_k$ , il faudra supprimer des unités de l'échantillon courant, plus précisément celles dont le nombre  $r_{hki}$  est le moins élevé, c'est-à-dire celles qui représentent les groupes de renouvellement les plus anciens dans l'échantillon. Pour que les unités du nouvel échantillon soient contenues dans l'intervalle fermé  $[0, f_h]$ , il est nécessaire de redéfinir  $r_{hki}$  comme  $r_{hki} - (f_k - f_h)$ , si  $r_{hki} \geq (f_k - f_h)$ , ou  $r_{hki} - (f_k - f_h) + 1$ , sinon. En supposant que les unités de population de la nouvelle strate  $h$  sont classées selon la valeur de  $r_{hki}$  en ordre croissant, posons  $b_{hi} = i / (M_h + 1)$ ,  $i = 1, 2, \dots, M_h$ . Nous pouvons déterminer de nouveaux numéros de renouvellement à l'aide des  $b_{hi}$ . Pour une nouvelle strate  $h$  donnée, posons  $N_h$  comme le nombre de groupes de renouvellement distincts. Créons  $N_h$  intervalles disjoints

$$I_n = \begin{cases} [(n - 1) / N_h, n / N_h] & \text{pour } n = 1, \dots, N_h - 1 \\ [(N_h - 1) / N_h, 1] & \text{pour } n = N_h. \end{cases}$$

L'union de ces intervalles correspond à l'intervalle fermé  $[0, 1]$ . Pour la nouvelle strate  $h$ , désignons les nouveaux numéros de renouvellement par  $D_1, D_2, \dots, D_{N_h}$  où  $D_{n_i} < D_{n_j}$  pour  $n_i < n_j$ ,  $N_h$ . L'unité  $i$  aura le numéro de renouvellement  $D_{n_i}$  si la valeur  $b_{hi}$  correspondante est comprise dans l'intervalle  $I_n$ . En supposant que l'on attribue de cette manière de nouveaux numéros de renouvellement aux  $M_h$  unités, les unités échantillonnées seront celles dont le numéro de renouvellement est compris dans l'intervalle  $[1, P_h]$ .

De fait, on opère le renouvellement en supprimant un groupe de l'échantillon et en lui en substituant un autre provenant de l'extérieur, le tout de façon modulaire.

Les entreprises naissantes peuvent être de deux types : lancement d'une entreprise ou changement d'activité qui fait qu'une entreprise auparavant exclue du champ de l'enquête en fait désormais partie. Les entreprises naissantes font l'objet d'une stratification et reçoivent un numéro de renouvellement qui est déterminé de la façon suivante. En supposant que le dernier numéro de renouvellement soit  $\ell$ , où  $1 \leq \ell \leq P$ , la  $q$ -ième entreprise naissante recevra le numéro de renouvellement  $(\ell + q) \bmod P$ . Ainsi, en supposant que  $b$  entreprises naissantes ont été assignées à des groupes de renouvellement, le dernier numéro de renouvellement utilisé dans les circonstances est  $(\ell + q) \bmod P$ . Autrement dit, la prochaine entreprise naissante recevra le numéro de renouvellement  $(\ell + q + 1) \bmod P$ . On peut ainsi connaître immédiatement le numéro de renouvellement grâce à la correspondance univoque entre le numéro d'attribution et le numéro de renouvellement.

Les entreprises mortes résultent soit de la cessation des activités commerciales, pour les unités faisant partie du champ de l'enquête, soit d'un changement d'activité qui fait qu'une entreprise auparavant incluse dans le champ de l'enquête en est désormais exclue. Les entreprises mortes qui font partie d'une strate à tirage complet sont immédiatement supprimées de l'échantillon et de la population. Celles qui font partie d'une strate à tirage partiel sont aussi immédiatement exclues de l'échantillon et de la population si elles sont identifiées comme des entreprises mortes par une source indépendante de l'enquête. Autrement, elles sont supprimées au bout d'un certain temps. Cet intervalle de temps doit être suffisamment long pour que puissent être détectées la plupart des entreprises mortes. Si jamais il reste des entreprises mortes dans l'échantillon ou la population, on leur impute une valeur nulle pour les besoins de l'estimation. De même, les valeurs de classification demeurent telles quelles jusqu'à ce qu'une source indépendante de l'enquête permette de dire qu'elles ont été modifiées.

## 2.2.4 Rééchantillonnage périodique

La base de sondage change continuellement à cause non seulement des créations et des disparitions d'entreprises, mais aussi des modifications que subissent les variables de classification servant à la stratification (c.-à-d. région géographique, branche d'activité et taille de cation servant à la stratification (c.-à-d. région géographique, branche d'activité et taille de l'entreprise). Le recours à l'estimation pour domaines (c.-à-d. à l'estimation pour sous-populations) traduit les changements qui peuvent survenir dans les variables de classification. Autrement dit, la plus récente classification sert à faire des totalisations à l'aide des poids d'échantillonnage originaux. Au bout d'un certain temps, il se peut que cette classification ait subi suffisamment de modifications pour nécessiter un examen d'un nouvel échantillon indépendant en tenant compte de ces modifications mais en ne s'occupant pas de l'échantillon courant. Une telle solution présente certains inconvénients du point de vue pratique. Le tirage d'un nouvel échantillon suppose i) qu'il faut intégrer les unités fraîchement échantillonnées dans l'échantillon de base, ii) que les règles qui déterminent le temps qu'une unité doit passer à l'intérieur et à l'extérieur de l'échantillon risquent de ne pas être respectées et iii) que les estimations risquent d'être modifiées de façon substantielle. C'est pourquoi il est souhaitable que le nouvel échantillon chevauche le plus possible l'échantillon courant. La méthode que nous décrivons ci-dessous vise justement à opérer un rééchantillonnage. C'est une version modifiée de la méthode de Kish et Scott (1971) qui repose sur la propriété que chaque groupe de renouvellement est un échantillon aléatoire simple tiré des groupes de renouvellement dans la population.

Au moment du rééchantillonnage, le processus de renouvellement n'est pas au même stade dans les différentes strates, d'où le fait que les intervalles d'échantillonnage n'ont pas la même limite inférieure ni la même limite supérieure. Ainsi, en supposant que le processus de renouvellement a débuté au temps  $t_1$  et que nous sommes actuellement au temps  $t_2$ , le nombre de renouvellements qui ont eu lieu depuis le début est  $r = t_2 - t_1 + 1$ . Au temps  $t_2$ , le ou les intervalles d'échantillonnage rattachés à une strate donnée  $k$  ( $k = 1, 2, \dots, K$ ) sont



chiffres 1 à  $P$ , tandis que la seconde ligne (dite "ligne de renouvellement") représente un classement aléatoire des éléments de la première ligne. Les numéros de groupe de renouvellement qui figurent sur la seconde ligne – car il s'agit bien de cela – permettent de déterminer quelles unités font partie de l'échantillon à un moment quelconque. Les  $M$  unités de la population sont assignées tour à tour aux groupes de renouvellement 1, 2, ...,  $P$ , la  $P$ -ième unité étant assignée au groupe de renouvellement  $P$ . La  $(P + 1)$ -ième unité est assignée au groupe de renouvellement 1, et ainsi de suite. Finalement, on obtient une série de groupes de renouvellement dont les " $\ell$ " premiers contiennent  $(a + 1)$  unités et les  $(P - \ell)$  autres renferment  $a$  unités. Le groupe auquel est assignée la  $M$ -ième unité est désigné comme le dernier groupe de renouvellement. On lui attribue le numéro  $\ell$  au moment de l'échantillonnage initial. Par la suite, les unités nouvellement créées sont assignées au groupe de renouvellement suivant, c.-à-d. au groupe  $\ell + 1$ .

Si  $M < P$ , les  $M$  unités de la population ne pourront être réparties qu'entre  $M$  des  $P$  groupes de renouvellement. Ces  $M$  groupes doivent être aussi équilibrés que possible si l'on veut obtenir la taille d'échantillon prévue,  $m = fM$ , à chaque passage de l'enquête. Dans les circonstances, on utilisera une matrice  $2 \times M$ . La première ligne est formée, dans l'ordre, des chiffres 1 à  $M$ . La seconde ligne est le résultat d'une randomisation de  $M$  nombres " $z_i$ ", où  $1 \leq z_i < z_j \leq P$  pour  $i \neq j$ ,  $i = 1, \dots, M$  et  $j = 1, \dots, M$ . Les nombres " $z_i$ " sont créés de la façon suivante.

- i) Trouver des entiers  $s$  et  $q$  qui satisfont l'équation  $P = sM + q$ , où  $q < M$  et  $s \geq 0$ .
- ii) Générer aléatoirement des nombres  $r_j$  ( $j = 1, \dots, M$ ) en supposant qu'ils peuvent prendre la valeur 0 ou 1, de telle sorte que  $q$  d'entre eux ont la valeur 1 et les autres  $(M - q)$  ont la valeur 0.
- iii) Choisir un entier aléatoire " $b$ " tel que  $1 \leq b \leq P$ .
- iv) Calculer  $z_1 = (b + r_1 - 1) \bmod P + 1$  et  $z_j = (z_{j-1} + s + r_j - 1) \bmod P + 1$  pour  $j = 2, \dots, M$ .
- v) Randomiser les nombres " $z_i$ ". Définir la suite des nombres " $z_i$ " randomisés comme  $z_{i1}, z_{i2}, \dots, z_{iM}$ .

On peut maintenant assigner les  $M$  unités de population, une à une, aux  $M$  groupes de renouvellement et, du même coup, leur associer un numéro de renouvellement. Le dernier groupe de renouvellement porte le numéro  $M$ . Pour ce qui est des futures nouvelles entreprises, elles seront assignées au groupe de renouvellement 1, 2, etc.

Le tirage et la mise à jour de l'échantillon sont devenues maintenant des opérations simples.

### 2.2.3 Tirage et mise à jour de l'échantillon

Au moment de l'échantillonnage initial, une strate donnée renfermera  $N = \min(M, P)$  groupes de renouvellement distincts. Les unités qui font partie de l'échantillon initial sont celles dont le numéro de renouvellement est inclus dans l'intervalle d'échantillonnage fermé  $[1, P]$ . Si  $M \geq P$ , le nombre  $n$  de groupes de renouvellement dans l'échantillon est égal à  $P$ . Si  $M < P$ , ce même nombre est approximativement égal à  $fN$  à cause de l'équilibrage.

On effectue le renouvellement de l'échantillon en déplaçant l'intervalle d'échantillonnage d'un groupe à chaque vague jusqu'à ce qu'on revienne au premier groupe. À la  $t$ -ième vague, les unités qui font partie de l'échantillon sont celles dont le numéro de renouvellement est compris dans l'intervalle suivant:

$$i) [(t - 1) \bmod P + 1, (t + P - 2) \bmod P + 1], \text{ si } (t - 1) \bmod P \leq (P - P)$$

et

$$ii) [1, (P - P) + (t - 1) \bmod P] \cup [(t - 1) \bmod P + 1, P], \text{ sinon.}$$



Le plan d'échantillonnage avec groupes de renouvellement présente plusieurs avantages par rapport aux deux plans mentionnés ci-dessus. En effet, avec le plan fondé sur des groupes de renouvellement, le nombre prévu d'unités dans chaque cycle de renouvellement est à peu près le même, ce qui n'est pas du tout le cas avec l'échantillonnage de Poisson. Par ailleurs, la suppression en bloc des unités disparues peut créer un déséquilibre entre les effectifs des divers groupes de renouvellement. On peut corriger la situation grâce à un rééchantillonnage périodique où l'on fait en sorte d'obtenir le plus haut degré de chevauchement possible et où on conserve le même mode de stratification et les mêmes fractions de sondage. Le renouvellement peut donc s'effectuer sans conséquence pour les unités, dans le respect des règles qui déterminent le temps qu'une unité doit passer à l'intérieur et à l'extérieur de l'échantillon. Cela pourrait ne pas être possible dans le cas de l'échantillonnage avec étalement à cause des légères modifications que subissent les nombres aléatoires par l'intermédiaire des créations et des disparitions d'entreprises. Ces effets peuvent devenir appréciables à longue échéance. Un autre avantage du plan fondé sur des groupes de renouvellement est qu'il permet la restructuration et la redéfinition des taux de sondage tout en conservant le plus haut degré de chevauchement possible.

## 2.2.1 Détermination du nombre de groupes de renouvellement

Supposons que pour une strate à tirage partiel donnée, le nombre d'unités de la population soit  $M$  et la fraction de sondage voulue,  $f$ . Soit  $t_m$  le nombre de vagues de consultation où une unité doit rester dans l'échantillon et  $t_{out}$ , le nombre minimum de vagues où une unité doit demeurer à l'extérieur de l'échantillon une fois qu'elle en a été supprimée. On calcule le nombre requis de groupes de renouvellement dans la population, " $P$ ", et de groupes de renouvellement dans l'échantillon " $p$ " de la façon suivante. Soit  $x = int[t_m (1 - f)/f + 0.5]$ , où  $int[\cdot]$  représente le plus grand entier de l'argument. Deux conditions se présentent:

- a) Si  $x \geq t_{out}$ , le nombre de groupes de renouvellement dans l'échantillon est  $p = t_m$  et le nombre de groupes de renouvellement dans la population est  $P = t_m + x$ .
- b) Si  $x < t_{out}$ , le nombre de groupes de renouvellement dans l'échantillon est

$$p = int \left[ \frac{1 - f}{f} t_{out} + 0.5 \right]$$

et le nombre de groupes de renouvellement dans la population est  $P = p + t_{out}$ .  
Il convient de souligner que le rapport  $p/P$  n'est qu'une approximation égale à  $f$  à cause de l'utilisation de nombres entiers.

## 2.2.2 Répartition des unités entre les groupes de renouvellement

Étant donné qu'au moment de l'échantillonnage initial,  $M$  unités de population doivent être réparties entre  $P$  groupes de renouvellement, nous pouvons nous trouver devant l'une ou l'autre de deux situations:  $M \geq P$  ou  $M < P$ .

Si  $M \geq P$ , chaque groupe de renouvellement pourra avoir au moins une unité. Supposons que  $M = aP + \ell$ , où  $a > 0$  et  $\ell \geq 0$  sont des entiers. Afin de former des groupes de renouvellement de taille comparable pour les besoins de l'échantillonnage initial et des tirages subséquents, nous procédons de la façon suivante. Nous nous servons d'une matrice  $2 \times P$  pour attribuer un numéro de renouvellement à chaque unité en gardant présent à l'esprit la nécessité de constituer des groupes de renouvellement de taille comparable. Cette matrice est utilisée à l'occasion de l'échantillonnage initial et pour l'ajout d'entreprises nouvellement créées. La première ligne de la matrice (dite "ligne d'attribution") est formée, dans l'ordre, des

strate primaire de sorte que les c.v. pour chaque groupe d'industries et chaque région géographique primaire et la taille de l'échantillon pour chaque strate primaire à l'aide d'une méthode itérative, de sorte que l'on obtient le c.v. global et les c.v. individuels voulus.

## 2.2 Plan d'échantillonnage

Les  $M$  unités de la population de chaque strate sont réparties aléatoirement en  $P$  groupes de renouvellement ( $P$  étant un nombre préalable) de sorte qu'initialement, les effectifs de deux groupes de renouvellement quelconques diffèrent d'au plus une unité. Le nombre de groupes de renouvellement est fonction des fractions de sondage et des règles qui déterminent le temps qu'une unité doit passer à l'intérieur et à l'extérieur de l'échantillon. Notons que l'on devra parfois entreprendre ces règles pour obtenir des estimations sans biais. Un échantillon aléatoire simple (EAS) de  $p$  groupes de renouvellement est prélevé parmi  $P$  groupes de renouvellement de la population. Le nombre  $p$  est déterminé de telle manière que le rapport  $p/P$  soit approximativement égal à la fraction de sondage  $f$  voulue. L'échantillon de l'enquête est formé de toutes les unités des  $p$  groupes de renouvellement échantillonnés. Son renouvellement s'opère par l'exclusion d'un groupe de renouvellement et l'introduction d'un autre. Les entreprises nouvellement créées sont réparties aléatoirement, une à une, entre les  $P$  groupes de renouvellement d'une manière systématique. Les entreprises disparues ne sont exclues des strates que si elles sont décelées à l'aide d'une source indépendante de l'enquête ou que si elles ont cessé leur activité depuis un certain temps. L'échantillon est mis à jour à l'aide des méthodes proposées par Kish et Scott (1971), qui prévoient une restructuration du fait que des unités de la population changent de strate. Le procédé de mise à jour est conçu de manière à modifier le moins possible l'échantillon. Cette méthode de "rééchantillonnage" présente des avantages indéniables. Premièrement, elle limite au maximum le nombre de nouvelles unités dans l'échantillon, ce qui est plus efficace du point de vue opérationnel, et elle permet de maintenir les coûts à leur minimum. Deuxièmement, elle limite le plus possible les risques de discontinuité auxquels sont exposées les estimations à cause du "rééchantillonnage".

D'autres plans d'échantillonnage peuvent servir à la sélection et au renouvellement de l'échantillon. Il s'agit de l'échantillonnage de Poisson et de l'échantillonnage avec échantillonnage (collocated sampling). Brewer, Early et Joyce (1972) et Sunter (1977) ont étudié les propriétés de ces plans d'échantillonnage. Selon la définition donnée par Hajék (1964), l'échantillonnage de Poisson fait que chaque unité de la population peut être prélevée de façon indépendante avec une probabilité donnée. La décision de prélever ou non une unité repose sur un tirage aléatoire indépendant ou tirage bernoullien. En supposant que la probabilité d'échantillonnage d'une unité donnée  $i$  est  $\pi_i$ , et qu'un nombre aléatoire  $u_i$  distribué uniformément dans l'intervalle  $(0,1)$  est généré, alors l'unité  $i$  sera prélevée si  $u_i \leq \pi_i$ . Cette probabilité d'échantillonnage correspond à la fraction de sondage de la strate à laquelle appartient l'unité. Bien que l'échantillonnage de Poisson se distingue principalement par la simplicité du mécanisme de renouvellement de l'échantillon, cette méthode n'est pas sans inconvénients. Sa principale faiblesse est que la taille effective de l'échantillon est une variable aléatoire. Cela peut poser des problèmes sérieux lorsqu'il y a peu d'unités dans une strate, éventualité qui peut conduire à des échantillons de taille nulle. Early et Brewer (1971) ont remédié au problème en recourant à l'échantillonnage avec échantillonnage. Cette méthode d'échantillonnage ressemble à la précédente à la différence qu'elle réduit la variation de la taille d'échantillon en répartissant également – au niveau de la cellule – les unités dans l'intervalle  $(0,1)$ . Brewer, Early et Hanif (1984) exposent en détail les propriétés de cette méthode. Alors que dans l'échantillonnage de Poisson, l'ajout des nouvelles entreprises et la suppression des entreprises disparues n'influent aucunement sur les nombres aléatoires qui se rattachent aux unités existantes, l'utilisation de l'échantillonnage avec échantillonnage aura inévitablement un effet – aussi modeste soit-il – sur ces nombres, ce qui pourrait perturber le plan de renouvellement par la violation des règles qui déterminent le temps qu'une unité doit passer à l'intérieur et à l'extérieur de l'échantillon.



Il est souhaitable que la somme des valeurs estimées des totaux de domaines soit égale au total de la population si les domaines sont exclusifs et exhaustifs. On peut réaliser cette condition en utilisant une série de poids qui est indépendante des domaines.

La section 2 sert à décrire le plan de sondage appliqué aux groupes de renouvellement. Dans la section suivante, nous présentons diverses méthodes d'estimation. La section 4 renferme les résultats d'une étude empirique visant à décrire le rendement de ces estimateurs dans diverses conditions d'enquête. Enfin, la section 5 présente les conclusions.

## 2. PLAN DE SONDAGE

### 2.1 Stratification et répartition de l'échantillon

La stratification d'un univers d'entreprises se fait habituellement selon un ou plusieurs des caractères suivants: branche d'activité, région géographique et taille de l'entreprise. La mesure de taille peut être unidimensionnelle (ex.: chiffre des ventes ou nombre d'employés) ou multidimensionnelle (ex.: revenu et actif). Dans le cas qui nous occupe, les strates primaires résultent du croisement des branches d'activité et des régions géographiques pour lesquelles on veut obtenir des estimations. À l'intérieur de ces strates, on retrouve des strates secondaires formées selon la mesure de taille des unités. Les strates secondaires sont constituées d'une strate à tirage complet et d'un certain nombre de strates à tirage partiel, qui font l'objet d'un échantillonnage. La strate à tirage complet est essentielle, compte tenu de la forte asymétrie de l'univers des entreprises. On peut déterminer la borne de la strate à tirage complet à l'aide d'une méthode proposée par Hidiroglou (1986). Cette méthode permet de calculer la borne optimale entre la strate à tirage complet et les strates à tirage partiel dans chaque strate primaire; cette borne est calculée de manière à minimiser la taille de l'échantillon global pour un coefficient de variation donné. Le calcul de cette borne tient compte aussi de ce que certaines unités seront des unités qui ont une structure complexe, par exemple une entreprise dont les activités font qu'elle peut être classée dans plus d'une strate primaire. Les bornes des strates à tirage partiel sont calculées soit au moyen de la règle de la racine de  $\int$  cumulative, proposée par Dalenius et Hodges (1959), soit au moyen de la règle de la racine de  $\sqrt{x}$  cumulative, de Hansen et coll. (1953). Dans le second cas,  $x$  est une variable de taille qui peut servir à la stratification de la population des entreprises.

La taille des échantillons pour les strates primaires est calculée de manière à respecter des niveaux de précision préétablis pour certaines estimations clés. Ce calcul doit aussi tenir compte de la manière dont les unités doivent être réparties entre les strates à tirage partiel. On suppose que l'information dont on se sert pour calculer la taille des échantillons est corrélée étroitement avec les variables clés préétablies. Une fois que les unités de l'échantillon à tirage complet sont prises en considération, on répartit le reste de l'échantillon entre les strates à tirage partiel dans chaque strate primaire proportionnellement à  $M^q$  ou à  $X^q$ , où  $M$  est le nombre d'unités dans la strate à tirage partiel et  $X$ , le total de la strate à tirage partiel pour la variable de taille étudiée. La puissance  $q$  ( $0 \leq q \leq 1$ ) est choisie en fonction du mode de répartition voulu. Par exemple, dans le cas d'une répartition proportionnelle à  $X^q$ , si l'on suppose que le rapport  $S_h/X_h$  ne varie pas sensiblement d'une strate à l'autre et que les facteurs de correction pour population finie peuvent être omis, le fait de poser  $q$  égal à 1 aboutit à une répartition de Neyman tandis que lorsque  $q$  tend vers zéro, les coefficients de variation dans les différentes strates tendent à s'équivaloir. Les avantages de la répartition à la puissance  $q$  sont analysés dans Bankier (1988). Il est possible de "corriger" la répartition de manière à obtenir la taille d'échantillon minimum voulue ou les poids maximum voulus pour chaque strate secondaire.

On peut associer les critères de fiabilité (c.-à-d. coefficients de variation) aux strates primaires de deux façons. Ou bien on définit un critère de fiabilité pour chaque strate primaire, ou bien, pour un coefficient de variation (c.v.) global (national) donné, on définit un c.v. pour chaque



# Méthodes d'échantillonnage et d'estimation pour des enquêtes infra-annuelles auprès des entreprises

M.A. HIDIROGLOU, G.H. CHOUDHRY et P. LAVALLÉE<sup>1</sup>

## RÉSUMÉ

Les auteurs proposent un plan de sondage particulier pour des enquêtes infra-annuelles auprès des entreprises; ce plan couvre les opérations suivantes: échantillonnage initial, renouvellement de l'échantillon et mise à jour de l'échantillon. Le plan d'échantillonnage consiste en un échantillonnage en grappes stratifié, la stratification étant faite selon la branche d'activité, la région géographique et la taille de l'entreprise. Le renouvellement de l'échantillon est assujéti à des règles précises concernant le temps qu'une unité doit passer dans l'échantillon ou à l'extérieur de l'échantillon. La mise à jour de l'échantillon concerne la sélection des entreprises naissantes (entreprises nouvellement créées), l'exclusion des entreprises mortes (entreprises ayant cessé leurs activités) et l'introduction de modifications dans les variables servant à la stratification, c.-à-d. branche d'activité, région géographique et taille de l'entreprise. Au moyen d'une étude empirique, les auteurs évaluent dans diverses conditions un certain nombre d'estimateurs – notamment, l'estimateur d'extension et l'estimateur par quotient sans biais de Mickey (1959) – pour les besoins de ce plan de sondage. Ils examinent aussi l'estimation de la variance au moyen de la méthode de linéarisation de Taylor et de la méthode du jackknife.

**MOTS CLÉS:** Enquêtes permanentes; mise à jour d'échantillon; estimateur par quotient; estimation de la variance.

## 1. INTRODUCTION

L'univers des enquêtes infra-annuelles auprès des entreprises évolue constamment en raison des créations et des disparitions d'entreprises, des fusions, des séparations et des changements de groupe. Le plan de sondage appliqué à un univers de ce genre doit avoir les caractéristiques suivantes. Premièrement, il doit produire des échantillons qui reflètent la structure changeante de la population. Deuxièmement, il doit permettre de répartir le fardeau de réponse en prévoyant le renouvellement de l'échantillon. Troisièmement, si des changements notables sont apportés au mode de stratification de l'univers, il doit permettre de tirer un nouvel échantillon qui reflète le changement de stratification et la modification probable des fractions de sondage. Le nouvel échantillon devrait recouvrir le plus possible l'ancien échantillon de manière à réduire au maximum le nombre de variations brusques dans les estimations et à contenir l'augmentation des coûts attribuable à l'introduction de nouvelles unités dans l'échantillon. Pour satisfaire à ces exigences, nous proposons un échantillonnage aléatoire simple de groupes de renouvellement (grappes) formés aléatoirement dans chaque strate. Chaque groupe de renouvellement est formé d'un ensemble d'unités ou d'une seule unité. Toutes les unités d'un groupe de renouvellement échantillonné font partie de l'échantillon. Le renouvellement de l'échantillon est assujéti à des conditions selon lesquelles les unités doivent demeurer dans l'échantillon pendant un certain temps et ne pas pouvoir y être réintroduites pendant une période minimum après qu'elles en ont été supprimées.

Pour des domaines d'intérêt donnés, nous définissons des estimateurs sans biais (ou quasiment sans biais) de même que les mesures de fiabilité correspondantes (coefficients de variation).

<sup>1</sup> M.A. Hidiroglou, G.H. Choudhry et P. Lavallée, Division des méthodes d'enquêtes-entreprises, Statistique Canada, 11<sup>ème</sup> étage, Immeuble R.H., Coats, Ottawa, Ontario, Canada, K1A 0T6.

MITOFSKY, W.J., et WAKSBERG, J. (1989). CBS models for election night estimation. Document présenté à l'American Statistical Association San Diego Winter Conference.

MURRY, J.P., LASTOVICKA, J.L., et BHALLA, G. (1989). Demographic and lifestyle selection error in mail-intercept data. *Journal of Advertising Research*, 46-52.

O'DAY, J., et WOLFE, A.C. (1984). Seat Belt Observations in Michigan – August/September 1983. Ann Arbor, Michigan: University of Michigan Transportation Research Institute.

SUDMAN, S. (1980). Improving the quality of shopping center sampling. *Journal of Marketing Research*, 17, 423-31.

STATISTIQUE CANADA (1979). *Méthodes de collecte et de diffusion des statistiques sur les voyages internationaux*. Section des voyages internationaux, Statistique Canada.

UNITED STATES TRAVEL AND TOURISM ADMINISTRATION (1989). In-flight Survey: Overseas and Mexican Visitors to the United States. Survey Period: January-December 1988. Washington, D.C.: United States Travel and Tourism Administration.

WELLS, J.K., WILLIAMS, A.F., et LUND, A.K. (1990). Seat belt use on interstate highways. *American Journal of Public Health*, 80, 741-742.

WOLFE, A.C. (1986). 1986 U.S. National Roadside Breathing Survey: Procedures and Results. Ann Arbor, Michigan: Mid-America Research Institute.

WOOD, D. (1978). The Eastleigh Carrefour: a hypermarket and its effects. London: U.K. Department of the Environment.

YATES, F. (1981). *Sampling Methods for Censuses and Surveys* (4ième éd.). London: Charles Griffin.

ZIEGLER, P.N. (1983). Guidelines for Conducting a Survey of the Use of Safety Belts and Child Safety Seats. Washington, D.C.: U.S. Department of Transportation.

Je désirerais exprimer ma reconnaissance aux nombreux chercheurs qui m'ont généreusement fourni des renseignements relatifs aux enquêtes sur les flux de populations humaines auxquelles ils ont participé.

## REMERCIEMENTS

## BIBLIOGRAPHIE

- BLAIR, E. (1983). Sampling issues in trade area maps drawn from shopper surveys. *Journal of Marketing*, 47, 98-106.
- BOWMAN, B.L., et ROUNDS, D.A. (1989). Restraint System Usage in the Traffic Population, 1988 Annual Report. Washington, D.C.: U.S. Department of Transportation.
- BRICK, M., et LAGO, J. (1988). The design and implementation of an observational safety belt use survey. *Journal of Safety Research*, 19, 87-98.
- BRYANT, E., et SHIMIZU, I. (1988). *Sample Design, Sampling Variance, and Estimation Procedures for the National Ambulatory Medical Care Survey*. Vital and Health Statistics, Series 2, No. 108, Washington D.C.: U.S. Government Printing Office.
- BUSH, A.J., et HAIR, J.F. (1985). An assessment of the mall intercept as a data collection method. *Journal of Marketing Research*, 22, 158-67.
- COCHRAN, W.G. (1977). *Sampling Techniques* (3ième éd.). New York: John Wiley.
- DOERING, Z.D., et BLACK, K.J. (1989). Visits to the National Air and Space Museum (NASM): Demographic Characteristics. Working Paper 89-1, Institutional Studies, Smithsonian Institution.
- DUPONT, T.D. (1987). Do frequent mall shoppers distort mall-intercept survey results? *Journal of Advertising Research*, 45-51.
- GATES, R., et SOLOMON, P.J. (1982). Research using the mall intercept: State of the art. *Journal of Advertising Research*, 22, 4, 43-49.
- GOODMAN, R., et KISH, L. (1950). Controlled selection - a technique in probability sampling. *Journal of the American Statistical Association*, 45, 350-372.
- GOUGH, J.H., et GHANGURDE, P.D. (1977). Survey of Canadian residents returning by land. *Techniques d'enquête*, 3, 215-231.
- GRIFFITHS, D., et ELLIOT, D. (1987). Sampling errors on the International Passenger Survey. Document non-publié, Social Survey Division, U.K. Office of Population Censuses and Surveys, London.
- HEADY, P. (1985). Note on some sampling methods for visitor surveys. *Survey Methodology Bulletin*. U.K. Office of Population Censuses and Surveys, 17, 10-17.
- HEBRINGA, S.G. (1985). The University of Michigan 1984 Library Cost Study: Final Report. Institute for Social Research, University of Michigan.
- HESS, I., RIEDEL, D.C., et FITZPATRICK, T.B. (1975). *Probability Sampling of Hospitals and Patients*. Ann Arbor, Michigan: Health Administration Press.
- JESSEN, R.J. (1978). *Statistical Survey Techniques*. New York: John Wiley.
- KISH, L., LOVEJOY, W., et RACKOW, P. (1961). A multi-stage probability sample for traffic surveys. *Proceedings of the Social Statistics Section, American Statistical Association*, 227-230.
- LEVY, M.R. (1983). The methodology and performance of election day polls. *Public Opinion Quarterly*, 47, 54-67.
- MITOFSKY, W.J. (1991). A short history of exit polls. Dans *Polling in Presidential Election Coverage*. (Eds. P. Lavrakas et J. Holley). Newbury Park, California: Sage.



durant environ de 5 à 6 minutes. Quand un intervieweur avait terminé une interview et que le répondant avait subi le test de l'éthyloscopie, l'intervieweur faisait signe au policier d'arrêter le prochain véhicule qui passait. Les interviews ont été réalisées pendant une période de deux heures à chaque emplacement sélectionné. On a compté le nombre de véhicules qui ont passé à l'emplacement dans la direction choisie pendant la période et le rapport entre ce nombre et le nombre d'interviews réalisées a été utilisé comme facteur de pondération dans l'analyse.

#### 4. CONCLUSIONS

Comme les exemples des sections précédentes le montrent, des considérations relatives au travail sur le terrain ainsi que le côté économique de la collecte des données jouent des rôles importants dans le choix du plan d'échantillonnage employé pour les enquêtes portant sur les personnes de passage. La longueur de l'intervalle de temps utilisé pour définir les UPE peut, par exemple, être dictée par la longueur d'un poste de travail approprié pour les travailleurs sur le terrain et cela peut entraîner l'utilisation d'UPE comportant des variations internes importantes dans le taux du flux. Par exemple, dans une enquête sur les passagers qui arrivent à une gare, le poste de travail d'un intervieweur du matin peut comprendre un flux en période de pointe quand les banlieusards arrivent tôt le matin et un flux moins important par la suite. Si l'on n'avait pas à faire correspondre les intervalles de temps des UPE au poste de travail des travailleurs sur le terrain, il serait préférable d'éviter une telle variation de flux à l'intérieur des UPE puisque cela entraîne des problèmes pour ce qui est de la façon de tirer un sous-échantillon dans les UPE sélectionnées.

Quand le flux des personnes dans une UPE est inégal, l'emploi de l'échantillonnage systématique, ou de tout plan d'échantillonnage EPSEM, pour sélectionner des personnes crée une charge de travail qui varie dans le temps. Si cette variabilité dans la charge de travail est importante, on a de la difficulté à décider comment affecter le personnel à l'UPE pour effectuer le travail sur le terrain, particulièrement dans le cas d'une enquête comportant des interviews par interrogation directe. L'affectation d'un nombre suffisant de personnes pour traiter les flux correspondants à des périodes de pointe n'est pas économique puisque les intervieweurs seront souvent inoccupés en dehors des périodes de pointe. Il arrive parfois qu'il soit préférable d'affecter un personnel suffisant pour traiter un flux un peu inférieur à celui que l'on rencontre pendant les périodes de pointe. Cela introduira une certaine non-réponse au moment où le flux correspond à une période de pointe parce qu'aucun intervieweur ne sera disponible pour réaliser une interview avec certaines personnes échantillonnées, mais en procédant de la sorte, on fait une meilleure utilisation du temps des intervieweurs.

L'utilisation la plus efficace du temps des intervieweurs consiste à leur faire interviewer la première personne qui arrive (ou qui part) après qu'ils ont terminé l'interview qu'ils sont en train de réaliser. Les plans de ce genre ont cependant le désavantage de ne pas produire d'échantillons probabilistes et, par conséquent, il y a un risque de biais dans les estimations de l'enquête. Quand on peut concevoir des plans d'échantillonnage probabilistes rentables, ces derniers doivent être préférés. Toutefois, le choix d'un plan d'échantillonnage dans lequel on choisit la première (ou la deuxième, ou la troisième) personne qui passe après qu'un intervieweur a terminé une interview est naturellement intéressant pour les enquêtes où les interviews sont réalisées par interrogation directe quand le flux de populations humaines est très variable et imprévisible. Quand ce genre de plan est employé, il est utile de compter le flux de personnes pendant de courts intervalles de temps. Ces totaux peuvent alors être utilisés pour faire des ajustements de pondération afin de compenser pour les probabilités de sélection inégales dues au flux irrégulier.

s'est arrêté brièvement, par exemple, à des feux de circulation. L'absence d'éclairage dans les rues peut empêcher l'observation de l'emploi des ceintures de sécurité la nuit à certains endroits. L'éthyloscopie exige que le véhicule soit arrêté et cela ne peut se faire de façon sécuritaire qu'en des emplacements où le véhicule a l'arrêt n'entraîne pas la circulation. Contrairement aux enquêtes par observation, les enquêtes entreprises dans le cadre desquelles on arrête les véhicules font face à un problème de non-réponse considérable.

Une méthode ingénieuse pour étudier l'emploi des ceintures de sécurité sur les routes entre les États est décrite par Wells et coll. (1990). Dans le cadre de cette étude, un observateur était assis derrière le conducteur d'une fourgonnette de tourisme qui se déplaçait à une vitesse inférieure à la vitesse moyenne de la circulation dans la voie de droite de la route. Depuis cette position, l'observateur pouvait déterminer si les occupants des sièges avant des automobiles, des camions légers et des fourgonnettes qui dépassaient la fourgonnette dans laquelle il prenait place dans la voie adjacente portaient leur ceinture diagonale.

Une méthode plus courante employée pour étudier le port des ceintures de sécurité consiste à faire des observations aux intersections des rues et aux sorties des autoroutes où l'on trouve des feux de circulation et parfois dans des centres d'achat et des terrains de stationnement (Ziegler 1983; Bowman et Rounds 1989). O'Day et Wolfe (1984) décrivent une enquête par observation de l'emploi des ceintures de sécurité au Michigan réalisée à l'aide de cette méthode. Ils ont échantillonné un certain nombre d'unités spatiales, sélectionné un certain nombre d'intersections avec feux de circulation dans ces unités, choisi des jours où les observations devaient être réalisées à ces intersections et échantillonné cinq périodes d'une heure chacune entre 08 h 00 et 20 h 00 pour réaliser les observations lors de chaque jour sélectionné. Pendant chacune des heures choisies, les observations ont été effectuées à une intersection différente. Les heures étaient sélectionnées selon un plan qui prévoit l'alternance d'une heure de travail et d'une heure de temps libre, les observateurs se déplaçant entre les intersections pendant les heures de temps libre. Les observations de l'utilisation des ceintures de sécurité ont été réalisées aux intersections sélectionnées aux moments précisés pour les véhicules qui s'étaient arrêtés aux feux de circulation. Quand plus d'un véhicule était arrêté, l'observation commençait avec le deuxième véhicule, à cause du biais associé au premier véhicule à s'arrêter à un feu de circulation. Afin d'obtenir des renseignements plus détaillés sur l'utilisation des dispositifs de protection pour enfants, on a aussi observé les véhicules qui entraient dans des centres commerciaux et des haltes routières.

La méthode habituelle utilisée pour analyser des données d'observation sur l'emploi des ceintures de sécurité consiste à calculer la proportion des personnes observées qui portaient leur ceinture de sécurité. Brick et Lago (1988) proposent une autre mesure, la proportion du temps estimé pendant lequel les occupants des sièges avant portent leur ceinture de sécurité dans les véhicules admissibles par rapport au temps total qu'ils passent dans ces véhicules. Pour leur enquête on a choisi un échantillon probabiliste de toutes les intersections des chaussées, qu'on y trouve ou non des feux de circulation. Afin d'éviter un biais dû à la sélection, on a dit aux observateurs l'endroit qu'ils devaient utiliser pour faire leurs observations et dans quelle direction ils devaient observer la circulation pendant la période précisée de 40 minutes au cours de laquelle ils devaient réaliser leurs observations. On a estimé le temps pendant lequel les occupants des véhicules étaient sur la route en divisant la longueur du segment de route menant à l'intersection par la vitesse moyenne estimée de la circulation sur ce segment de route. Ce temps estimé a été utilisé comme facteur de pondération dans l'analyse.

Les considérations en matière d'échantillonnage pour les enquêtes au bord de la route au cours desquelles on a recours à l'éthyloscopie sont généralement semblables à celles qui s'appliquent aux enquêtes sur l'utilisation de la ceinture de sécurité, sauf que les endroits où les données peuvent être recueillies doivent être des endroits où il est possible d'arrêter les véhicules sans danger. Au cours de la National Roadside Breathtesting Survey de 1986 aux É.-U., les agents de police locaux ont collaboré à l'enquête en demandant aux conducteurs sélectionnés de s'arrêter et en les orientant vers les intervieweurs chargés de réaliser l'enquête (Wolfe 1986). Les interviews



heures d'ouverture. La deuxième étape consistait à compter les sorties des groupes de personnes qui font leurs courses dans les magasins échantillonnés pendant les heures sélectionnées, le comptage étant effectué pendant quinze minutes de chaque période d'une heure. L'opération de comptage a été effectuée pendant un mois. En fonction des totaux obtenus, les interviews ont été réparties entre les genres de magasin et les jours de la semaine et à des magasins et des heures particulières. On a ensuite demandé aux intervieweurs d'interviewer le nombre précisé de personnes quittant le magasin, en interviewant la personne qui quittait le magasin après qu'ils avaient terminé une interview. L'échantillon porte sur les visiteurs dans un magasin et les personnes qui font leurs courses pouvaient visiter plusieurs magasins lors d'un déplacement en particulier au centre commercial. On a demandé aux répondants s'ils avaient déjà visité des magasins du centre commercial lors de ce déplacement particulier et aussi combien d'autres magasins ils comptaient visiter. Ces données ont été utilisées pour élaborer des poids afin d'effectuer des analyses des déplacements.

Le second type d'enquêtes réalisées dans les centres commerciaux utilise les personnes sélectionnées aux centres commerciaux comme un échantillon à l'aveuglette de la population dans son ensemble. Les interviews sur le vif de ce genre sont couramment utilisées dans les études de marché (Bush et Hair 1985; Gates et Solomon 1982). Les procédures manquent souvent de contrôle et les échantillons pourraient bien être biaisés. Les questions en jeu sont examinées par Sudman (1980), quand il se penche sur les procédures à suivre pour échantillonner les centres commerciaux, les emplacements dans les centres sélectionnés et les périodes de temps afin d'améliorer les plans d'échantillonnage, et par Blair (1983), Dupont (1987) et Murry et coll. (1989).

### 3.7 Enquêtes sur la circulation routière

Une forme d'enquête sur la circulation routière porte sur la circulation qui passe par un ou plusieurs emplacements. On peut appliquer, de façon relativement simple, des plans d'échantillonnage temporels et dans l'espace à de telles enquêtes. Kish et coll. (1961), par exemple, décrivent le plan d'échantillonnage pour une enquête origine-destination des véhicules qui utilisent les ponts et les tunnels de la Port of New York Authority au-dessus et au-dessous du fleuve Hudson, respectivement, pendant l'année 1959. Un plan d'échantillonnage stratifié à quatre degrés avec PPTB a été utilisé pour cette enquête. Les UPB étaient des combinaisons de postes de travail de huit heures et de ponts ou de tunnels particuliers. Un échantillon de ces UPB a été choisi au cours de la première étape, un échantillon de voies de circulation menant à des postes de péage contigus (les emplacements) a été choisi lors de la deuxième étape à l'intérieur des UPB sélectionnées, un échantillon de voies particulières a été choisi lors de la troisième étape dans les emplacements retenus et, finalement, un échantillon systématique de véhicules a été choisi dans les voies sélectionnées. Les intervieweurs restaient à un emplacement échantillonné pendant quatre heures et se déplaçaient chaque heure d'une voie de circulation à l'autre selon un plan préalable.

Un autre type d'enquête sur la circulation routière se rapporte à la circulation générale sur la route. Les enquêtes sur les passagers des voitures de tourisme pour étudier l'utilisation des ceintures de sécurité et la concentration d'alcool dans le sang des conducteurs sont de ce type. Une discussion complète des questions complexes relatives au plan d'échantillonnage utilisé dans ces enquêtes dépasse la portée du présent article; on se limitera plutôt à quelques observations générales.

La méthode de collecte des données à utiliser exerce une forte influence sur les procédures d'échantillonnage employées pour une enquête générale sur la circulation. Le port de la ceinture de sécurité est surtout étudié à l'aide de méthodes d'observation, alors que la mesure de la concentration d'alcool dans le sang fait habituellement appel à l'éthyloscopie. L'utilisation de la ceinture diagonale par les personnes qui occupent les sièges avant peut être observée dans la circulation en mouvement, mais l'emploi des ceintures ventrales et le port des ceintures de sécurité par les autres passagers d'un véhicule ne peuvent être observés que lorsque ce dernier



d'échantillonnage semblables ont été utilisés pour les ports d'entrée terrestres et pour les aéroports et, par conséquent, nous ne décrivons ici que le plan d'échantillonnage utilisé pour les ports d'entrée terrestres.

À un certain moment, le plan d'échantillonnage employé pour les ports d'entrée terrestres, dans le cas des résidents du Canada revenant au pays qui avaient passé au moins une nuit à l'étranger, consistait à distribuer des questionnaires d'enquête à tout groupe de voyageurs à chaque quatrième jour pendant toute l'année, les jours étant choisis par échantillonnage systématique. Ce plan s'est révélé impraticable parce qu'il arrivait trop souvent que les agents des douanes ne l'appliquaient pas correctement. Il a donc été remplacé par un plan basé sur une période de sondage dans le cadre duquel on a attribué à un port d'entrée terrestre deux périodes de sondage pour chaque trimestre de l'année, pendant lesquelles les questionnaires devaient être distribués. On prévoyait que ces périodes de sondage devaient durer de six à dix jours, avec des périodes de sondage successives commençant à des intervalles d'environ 6 1/2 semaines (Gough and Ghangurde 1977). Le nombre de questionnaires envoyés à un port d'entrée terrestre pour une période de sondage particulière était déterminé à partir de la circulation prévue à ce port d'entrée. On a ensuite demandé aux agents des douanes de commencer la distribution des questionnaires une journée donnée et de continuer à les distribuer jusqu'à ce qu'il ne leur en reste plus. Ce plan d'échantillonnage est adapté à des restrictions opérationnelles découlant de l'emploi des agents des douanes, pour lesquels l'enquête n'est qu'une préoccupation secondaire, comme travailleurs sur le terrain pour l'enquête. Le plan d'échantillonnage à certains désavantages importants, mais le fait que le taux de réponse soit de 20% ou moins constitue peut-être une préoccupation plus sérieuse.

Dans les enquêtes sur les voyageurs internationaux réalisées aux E.-U. et au Canada on s'en remet à la collaboration d'autres organismes pour effectuer le travail sur le terrain. Cette collaboration présente des avantages remarquables au niveau des coûts, mais il faut sacrifier la possibilité d'appliquer des contrôles rigoureux sur les procédures de travail sur le terrain. Les enquêtes sur les voyageurs se déplaçant par voie aérienne ou maritime réalisées au R.-U. emploient des procédures d'interviews par interrogation directe plus coûteuses. La International Passenger Survey réalisée au R.-U. en 1984 incluait, comme strates, les trois aéroports de Heathrow ainsi que les aéroports de Gatwick et de Manchester (Griffiths et Elliot 1987). Dans chaque aéroport, les jours ont été divisés en matinées et en après-midi et ces périodes constituaient les UPE. Un échantillon stratifié d'UPE a été sélectionné et des échantillons systématiques de passagers ont été choisis dans les UPE sélectionnées. Un échantillon d'UPE pour d'autres aéroports a aussi été inclus. Deux procédures différentes de collecte des données étaient utilisées dans les ports de mer. Dans certains ports de mer, les intervieweurs voyageaient sur le bateau, interviewant les passagers pendant la traversée. Dans le premier cas, les intervieweurs travaillaient par postes qui comprenaient plusieurs départs et le poste devenait l'UPE. Dans le dernier cas, les traversées étaient les UPE.

### 3.6 Enquêtes réalisées dans les centres commerciaux

Deux types d'enquêtes sont réalisées dans les centres commerciaux. L'un de ces types d'enquêtes vise à décrire les caractéristiques socio-économiques des personnes qui font leurs courses, la région où elles habitent ainsi que leurs activités de magasinage au centre commercial. Dans l'autre type d'enquêtes on utilise le centre commercial comme endroit commode pour obtenir des échantillons de personnes provenant de l'ensemble de la population de la région. Un exemple d'une enquête du premier type est une étude qui a été réalisée afin d'étudier l'incidence de l'ouverture d'un hypermarché dans les faubourgs de la ville de Southampton en Angleterre (Wood 1978). Des enquêtes auprès des personnes qui font leurs courses ont été réalisées dans quatre centres commerciaux avoisinants avant et après l'ouverture de l'hypermarché (et aussi dans l'hypermarché lui-même). Dans chacun de ces centres, la première étape du processus d'enquête était le dénombrement de tous les points de vente ainsi que de leurs

la population. Dans la deuxième étape, les médecins sont échantillonnés à partir de listes dans les UPB sélectionnées avec des intervalles d'échantillonnage différents d'une UPB à l'autre afin de tenir compte des probabilités inégales de sélection des UPB (dans les enquêtes plus récentes, différentes classes de spécialités sont échantillonnées à des taux différents). Les médecins échantillonnés sont ensuite répartis au hasard, d'une façon équilibrée, dans une des 52 semaines de déclaration de l'année. On demande à chaque médecin de relever des renseignements pour un échantillon systématique des visites de ses patients qui se produisent pendant la semaine échantillonnée; l'intervalle d'échantillonnage est choisi afin de donner approximativement 30 visites échantillonnées au cours de la semaine. Un intervalle d'échantillonnage de 1, 2, 3 ou 5 est choisi pour un médecin donné en fonction du nombre de visites au cabinet que le médecin s'attend à recevoir pendant la semaine et du nombre de jours où il ou elle prévoit recevoir des patients. Les procédures de travail sur le terrain consistent à tenir un journal de l'arrivée des patients à des fins d'échantillonnage, puis à remplir un bref relevé de seize éléments pour chaque visite échantillonnée.

La NAMCS est une enquête portant sur les visites des patients et non sur les patients. À ce titre, elle fournit des renseignements utiles sur la nature du travail des médecins en fonction des visites – la fréquence d'utilisation des tests de diagnostic, les thérapies fournies et les caractéristiques démographiques des patients reçus. Toutefois, l'enquête ne fournit pas d'estimations en fonction d'un patient, comme les traitements et les résultats pour les périodes où les patients sont malades.

### 3.5 Enquêtes sur les passagers internationaux

Un certain nombre de pays réalisent des enquêtes sur leurs voyageurs internationaux, tant les personnes qui entrent dans leur pays que celles qui le quittent, par voie terrestre, maritime ou aérienne. Dans la présente sous-section nous décrivons brièvement les plans d'échantillonnage utilisés pour une enquête sur les passagers internationaux qui se déplacent par voie aérienne réalisée par les États-Unis, pour des enquêtes sur les passagers internationaux qui voyagent par voie aérienne ou terrestre réalisées par le Canada et pour une enquête sur les passagers internationaux qui se déplacent par voie aérienne ou maritime réalisée par le R.-U.

La United States Travel and Tourism Administration réalise une In-flight Survey of International Air Travelers pour enquêter tant sur les voyageurs étrangers aux États-Unis que sur les résidents des États-Unis qui voyagent à l'étranger (voir, par exemple, United States Travel and Tourism Administration 1989). L'enquête est réalisée avec la collaboration volontaire d'environ trente sociétés aériennes. Un échantillon stratifié des vols réguliers est sélectionné pour la troisième semaine de chaque mois et tous les passagers de ces vols sont inclus dans l'échantillon. On fournit aux sociétés aériennes qui participent à l'enquête une trousse d'enquête qui renferme des instructions et des questionnaires dans les langues appropriées pour chaque vol échantillonné. Le personnel de cabine de la société aérienne distribue, dans les airs d'embarquement ou au cours des envolés, des questionnaires à remplir soi-même à tous les passagers adultes et il les recueille avant le débarquement. La non-réponse constitue un problème grave pour ces enquêtes. Dans le cas de l'enquête de 1988 sur les visiteurs aux États-Unis, aucun questionnaire n'a été retourné de la moitié des troussees remises au cours de ces envolés. Dans le cas des envolés pour les personnes qui ne résident pas aux États-Unis était de 44% et pour les résidents des E.-U., ce taux n'était que de 20%.

La Section des voyages internationaux de Statistique Canada réalise des enquêtes sur les voyages internationaux aux aéroports ainsi qu'aux ports d'entrée terrestres du Canada. Les enquêtes sont entreprises en collaboration avec Revenu Canada – Douanes et Accise, et ce sont les agents des douanes qui sont responsables de distribuer les questionnaires à remplir soi-même et à retourner par la poste. L'exposé présenté ici est basé sur le rapport produit par la Section des voyages internationaux, Statistique Canada (1979). Il reflète les plans d'échantillonnage qui s'appliquaient avant certains changements qui ont été apportés récemment. Des plans



La distinction entre la "visite" et le "visiteur" est particulièrement saillante pour cette enquête. Les personnes pouvaient, bien entendu, visiter le musée pendant plusieurs jours au cours de la période de l'enquête et elles pouvaient aussi visiter le musée plusieurs fois pendant la même journée. Cette dernière possibilité est particulièrement probable dans le cas du National Air and Space Museum parce que l'entrée au musée est gratuite et, par conséquent, rien n'incite les visiteurs à entrer une seule fois. Compte tenu de cette situation, il pourrait être approprié de définir les entrées multiples au cours d'une même journée comme une seule visite pour certains types d'analyses. Dans certains cas, cette définition pourrait être appliquée en limitant l'analyse aux personnes qui sortent du musée pour la première fois le jour échantillonné.

### 3.3 Sondages des votants

Un certain nombre des principales agences de presse réalisent des sondages auprès des électeurs les jours d'élection aux Etats-Unis (Levy 1983; Mitofsky 1991). Les électeurs sont échantillonnés au moment où ils quittent les bureaux de scrutin. On demande aux personnes sélectionnées de remplir un questionnaire bref et simple et de le remettre dans une boîte de scrutin. Un questionnaire typique comprend environ 25 questions dans lesquelles on demande comment le répondant a voté, quelle est sa position sur des questions clés, quelle est l'opinion du répondant sur divers sujets et quelles sont ses caractéristiques démographiques. Les taux de refus, pour les sondages des votants réalisés par la société CBS se sont établis, en moyenne, à 25% lors des dernières élections (Mitofsky et Waksberg 1989).

L'échantillonnage des électeurs pour les enquêtes électorales emploie habituellement un plan d'échantillonnage simple à deux degrés. Lors de la première étape, on tire un échantillon stratifié avec PPTB des circonscriptions électorales, où la mesure utilisée pour la taille est le nombre d'électeurs dans la circonscription. Lors de la deuxième étape, on choisit un échantillon systématique des électeurs quittant le bureau de scrutin, avec un intervalle d'échantillonnage choisi pour produire un échantillon approximativement EPSEM des électeurs dans les Etats. Habituellement, un seul intervieweur est affecté à chaque circonscription sélectionnée. Le travail sur le terrain est simple quand le bureau de scrutin a une seule sortie et qu'on permet à l'intervieweur de s'en approcher. Quand il y a deux sorties ou plus, les intervieweurs vont d'une sortie à l'autre, travaillant à chacune d'entre elles pour des périodes de temps déterminées. Quand cela se produit, l'intervalle d'échantillonnage doit être modifié en conséquence. Dans certains Etats, on ne permet pas aux intervieweurs de s'approcher à moins d'une certaine distance des bureaux de scrutin et cela peut créer des problèmes si les électeurs ont alors la possibilité de partir dans différentes directions avant que l'intervieweur ne réussisse à communiquer avec eux.

### 3.4 Enquête sur les soins médicaux ambulatoires

La National Ambulatory Medical Care Survey (NAMCS) des Etats-Unis emploie un plan d'enquête basé sur les flux de populations humaines pour recueillir des données sur les consultations médicales dans le cas des médecins en pratique privée qui dirigent les soins aux patients (Bryant et Shimizu 1988). La NAMCS a été réalisée un certain nombre de fois depuis qu'elle a été mise en oeuvre en 1973. Pour chaque enquête, la collecte des données a été répartie pendant toute l'année civile de l'enquête afin de fournir des estimations annuelles des caractéristiques des visites. On a toutefois demandé à chacun des médecins échantillonnés de fournir des renseignements pour un échantillon des visites qu'il reçoit pendant une seule semaine. On obtient une couverture annuelle en demandant à différents médecins échantillonnés de produire une déclaration pour différentes semaines de l'année.

L'échantillon de la NAMCS est basé sur un plan d'échantillonnage complexe à trois degrés qui a varié dans le temps. Un aperçu sommaire du plan d'échantillonnage suffira pour les besoins actuels; pour plus de détails, le lecteur est prié de se reporter à Bryant et Shimizu (1988). La première étape du plan d'échantillonnage de la NAMCS est le choix d'un échantillon stratifié avec PPTB d'UPB aréolaires, sélectionné avec une probabilité proportionnelle à la taille de



Les UPF ont été sélectionnées par échantillonnage avec PPTF, où la taille estimée pour une UPF était le nombre estimé de personnes sortant de cette bibliothèque dans la période de temps précisée. Des estimations grossières de ces nombres ont été obtenues à partir de la fréquentation quotidienne moyenne en novembre 1983, selon les chiffres enregistrés par les tourniquets quand ils étaient disponibles et des estimations fournies par les bibliothécaires quand ce n'était pas le cas et en fonction d'une hypothèse selon laquelle le volume des sorties des bibliothèques était deux fois plus élevé entre 9 h 30 et 17 h 30 qu'en d'autres temps. Les bibliothèques ont été stratifiées en quatre types et, dans chaque strate, on a utilisé une sélection contrôlée pour obtenir une distribution proportionnelle de l'échantillon parmi les bibliothèques, les jours de la semaine et les intervalles de temps.

Dans chaque UPF sélectionnée, on a choisi pour l'enquête un échantillon systématique de personnes sortant de la bibliothèque, avec l'intervalle d'échantillonnage déterminé afin de donner un échantillon global EPSEM des visites. On a fourni aux travailleurs sur le terrain une feuille d'inscription sur laquelle figuraient les entiers de 1 à 430 et où les numéros sélectionnés étaient marqués. Tout ce que les travailleurs avaient à faire était de cocher un numéro pour chaque personne sortant de la bibliothèque et de choisir les personnes associées aux numéros échantillonnés. Cette méthode présente l'avantage que les pas d'échantillonnage fractionnaires sont faciles à traiter. Quand on prévoyait que le volume des sorties pour une UPF échantillonnée devait être faible, un seul travailleur devait faire le comptage et communiquer avec les personnes échantillonnées. Quand le volume des sorties était élevé, le travail était réparti entre deux travailleurs, un qui effectuait le comptage et l'autre qui communiquait avec les personnes échantillonnées. Il fallait aussi employer plus d'un travailleur dans les bibliothèques disposant de plus d'une sortie.

### 3.2 Une enquête sur les visites dans un musée

Une enquête par interrogation directe des visiteurs sortant du National Air and Space Museum à Washington, D.C. a été réalisée de la mi-juillet jusqu'en décembre 1988 (Doering et Black 1989). L'interview, d'une durée d'environ quatre à six minutes, visait à recueillir des données sur les antécédents socio-démographiques de la personne échantillonnée, sur son lieu de résidence, sur ses activités lors de la visite, sur les objets exposés qui l'ont particulièrement intéressée, sur la raison de sa visite, sur la taille et le type de groupe, si elle faisait partie d'une visite en groupe et sur le mode de transport utilisé. Les enfants de moins de 12 ans et les personnes travaillant au musée étaient exclus de l'enquête. Les données ont été recueillies auprès de 5,574 répondants, avec un taux de réponse de 86%.

Chaque jour de la période d'enquête était divisé en deux demi-journées. Les interviews étaient réalisés pendant l'une de ces demi-journées tous les deux jours, avec alternance entre les matinées et les après-midis. Pendant l'été, le public pouvait utiliser trois sorties du musée, alors que plus tard au cours de l'année seulement deux d'entre elles étaient ouvertes. Pendant les demi-journées sélectionnées, la collecte des données d'enquête se faisait par rotation, sur une base horaire, entre les sorties qui étaient ouvertes. L'équipe de travailleurs sur le terrain affectée à une sortie à une heure échantillonnée était composée d'un ou de deux compteurs et de deux intervieweurs. Le chef compteur utilisait un compteur mécanique ainsi qu'un chronomètre pour suivre le nombre de personnes sortant du musée et pour tenir un registre qui donnait le nombre de personnes sortant dans chaque intervalle de dix minutes pendant l'heure. Le chef compteur identifiait aussi les personnes à interviewer. Le choix des personnes échantillonnées était fait pour que les intervieweurs ne soient jamais inoccupés. Le chef compteur remarquait quand un intervieweur avait terminé une interview et était prêt à en commencer une autre et il choisissait alors la cinquième personne sortant après ce moment comme la prochaine personne échantillonnée. Les comptes des flux de personnes pendant une période de dix minutes étaient utilisés dans l'analyse pour élaborer des poids afin de compenser pour la variation dans la sélection aléatoire associée au flux variable de personnes dans le temps.

à propos de la façon dont l'échantillon a été obtenu, une hypothèse courante étant que tous les éléments dans la population ont une chance égale d'être choisis. Quand les hypothèses ne se vérifient pas, cela peut entraîner un biais considérable dans les estimations de l'enquête. Les visiteurs peuvent être échantillonnés soit au moment où ils entrent dans un endroit, soit au moment où ils le quittent. Si l'on recherche des données sur les activités des visiteurs dans cet endroit et sur leurs opinions à propos de cet endroit, l'échantillon doit être composé seulement des personnes qui quittent l'endroit. Dans d'autres cas, le choix entre le fait d'échantillonner les personnes qui entrent dans un endroit ou qui en sortent peut dépendre de la nature des flux de populations humaines. Il peut, par exemple, être difficile d'échantillonner et d'interviewer des personnes qui quittent un cinéma ou une salle de théâtre parce que tous les spectateurs quittent les lieux en masse et qu'ils ne voudront pas être retardés. Par contre, ces personnes peuvent être échantillonnées et interviewées facilement quand elles font la queue pour entrer dans le cinéma ou la salle de théâtre.

En concluant cette section, il faudrait attirer l'attention sur le fait que les échantillons décrits ici sont des échantillons de visites et non de visiteurs. Le plan d'échantillonnage standard à deux degrés peut produire un échantillon EPSEM de visites, mais ce n'est pas la même chose qu'un échantillon EPSEM de visiteurs à moins que chaque visiteur ne visite l'endroit étudié (l'exposition de sculptures) qu'une fois (ou que tous les visiteurs visitent l'endroit étudié le même nombre de fois). Pour la majorité des enquêtes sur les flux de populations humaines, la visite, plutôt que le visiteur, est l'unité d'analyse appropriée. Il y a, toutefois, des situations où l'unité d'analyse est problématique. Si l'on utilise la visite comme unité d'analyse, le chercheur pourrait facilement accepter des visites à l'exposition de sculptures lors de deux jours différents comme des visites distinctes, mais il pourrait ne pas accepter de traiter deux entrées la même journée (une, peut-être, après être sorti pendant une brève période pour des rafraîchissements) comme deux visites. L'utilisation du visiteur comme unité d'analyse présente des problèmes graves à cause de la question des visites multiples et du fait que les visiteurs ne pourront pas déclarer leurs visites multiples. Il se peut qu'ils puissent se souvenir, assez bien, de leurs visites antérieures, mais habituellement ils ne pourront prévoir, avec précision, leurs visites futures.

### 3. EXEMPLES

La présente section renferme certains exemples d'enquêtes sur les flux de populations humaines afin de montrer la gamme étendue d'applications et pour illustrer certaines des considérations spéciales inhérentes à des situations particulières.

#### 3.1 Une enquête sur l'utilisation des bibliothèques

Une enquête sur l'utilisation des 18 bibliothèques de la University of Michigan a été réalisée en 1984 (Heeringa 1985). On a demandé à chaque personne échantillonnée sortant d'une bibliothèque si elle avait utilisé les documents et les services de la bibliothèque pendant cette visite. Dans l'affirmative, on demandait à la personne de remplir un bref questionnaire de sept questions portant sur les documents consultés et sur les services utilisés. La majorité des 5,184 répondants ont rempli le questionnaire sur place et l'ont remis aux travailleurs réalisant l'enquête sur le terrain; d'autres les ont retournés par l'intermédiaire du service de messageries de l'université. Un taux de réponse de 96% a été obtenu.

Le plan d'échantillonnage suivait le plan d'échantillonnage à deux degrés selon le temps/l'endroit décrit dans la section 2. L'enquête visait toute l'année civile 1984. Chaque jour où les bibliothèques étaient ouvertes a été divisé en dix intervalles de temps de deux heures, de 7 h 30 jusqu'à 3 h 30 le matin du jour suivant, l'intervalle de deux heures étant choisi parce qu'il constituait un poste de travail approprié pour les personnes réalisant l'enquête sur le terrain. Les LPE ont alors été définies comme des combinaisons d'intervalles de temps/de bibliothèques.



la prochaine personne à interviewer soit choisie. Si le flux est irrégulier, on doit prendre les dispositions nécessaires pour traiter les périodes de pointe (par exemple, l'arrivée, à l'exposition de sculptures, d'un autocar rempli de visiteurs).

La sélection avec PPTÉ des UPÉ permet de rendre égale la taille des sous-échantillons pour chaque UPÉ échantillonnée. Dans les enquêtes réalisées par interrogation directe, la charge de travail de l'intervieweur est ainsi à peu près la même pour chaque UPÉ sélectionnée et, par conséquent, la taille de l'équipe d'intervieweurs affectée à chaque UPÉ peut être la même. Il se produit toutefois un problème quand la mesure de la PPTÉ utilisée pour sélectionner l'UPÉ lors de la première étape est entachée d'une erreur considérable. Par exemple, un orage peut réduire substantiellement le nombre de visiteurs de l'exposition de sculptures un samedi après-midi en particulier ou, un congé non prévu peut augmenter considérablement le nombre de visiteurs un autre jour. Dans le premier cas, le fait d'appliquer dans cette UPÉ un intervalle d'échantillonnage inversement proportionnel à sa taille estimée laissera les intervieweurs très souvent inoccupés; alors que, dans le second cas, cela entraînera une charge de travail que les intervieweurs ne pourront pas traiter. Une modification qui peut être adoptée dans de tels cas consiste à changer l'intervalle d'échantillonnage au début de la collecte des données pour employer un intervalle qui est plus approprié au flux réel des visiteurs. Puisque cette modification détruit la propriété EPSEM de l'échantillon, il faut utiliser des poids dans l'analyse de l'enquête.

Une restriction générale qui s'applique à l'échantillonnage systématique des visiteurs pour des UPÉ sélectionnées est que, si la longueur de l'intervalle d'échantillonnage est rendue suffisante pour permettre aux intervieweurs de traiter les flux lors des périodes de pointe, ces derniers passent une bonne partie de leur temps sans travail. Par contre, si l'intervalle d'échantillonnage est réduit, les intervieweurs sont plus occupés, mais ils ne peuvent traiter les flux lors des périodes de pointe. Diverses méthodes ont été proposées pour surmonter ces problèmes (Heady 1985). Une méthode consiste à prendre un échantillon systématique d'intervalles de temps (disons à toutes les 10 minutes) et de choisir le prochain visiteur qui entre après chaque intervalle de temps échantillonné. Cette méthode pourrait présenter un certain intérêt sur le terrain, mais elle ne produit pas un échantillon probabiliste des visiteurs. La probabilité d'être choisie des personnes qui arrivent lors des périodes de grande activité est moindre, comme c'est le cas pour les personnes qui voyagent en groupe et les habitudes de déplacement de ces dernières peuvent avoir un effet inconnu sur la probabilité qu'ont ces personnes d'être sélectionnées. L'échantillon produit par cette procédure n'est certainement pas un échantillon EPSEM. On peut tenter de compenser le biais dû à la sélection qui défavorise les visiteurs arrivant pendant les périodes de grande activité en divisant l'intervalle de temps pour des UPÉ sélectionnées en un ensemble d'intervalles beaucoup plus courts et en tenant un journal des arrivées dans chacun de ces intervalles. On peut alors employer la pondération pour compenser la variation dans le flux pendant les intervalles plus courts.

Une autre méthode qui peut être employée à la place de l'échantillonnage systématique des visiteurs consiste à choisir la prochaine personne à entrer (ou à sortir) après la fin de la dernière interview. Dans cette méthode, les premières personnes qui arrivent après des intervalles dans le flux de visiteurs, peut-être les chefs de groupes, ont évidemment des chances plus élevées d'être choisies. Il se peut aussi que les intervieweurs accélèrent ou ralentissent délibérément l'interview qu'ils sont à réaliser afin d'éviter ou de choisir une personne en particulier. Pour ces raisons, on a employé des variantes de cette méthode dans le cadre desquelles on choisit la  $n^{\text{e}}$  personne qui entre ou qui sort après la fin d'une interview, où  $n$  peut avoir comme valeur 2, 3, 4, ou 5. Ces méthodes qui peuvent être utilisées à la place d'un échantillonnage systématique simple des visiteurs utilisent plus efficacement le temps des intervieweurs et permettent donc d'obtenir des échantillons plus grands avec un budget donné pour le terrain. Toutefois, elles produisent des échantillons non probabilistes, avec le risque de biais de sélection que ce genre d'échantillonnage suppose. L'échantillonnage probabiliste fournit la sécurité de l'inférence statistique objective sans que l'on ait à faire des hypothèses à propos du processus de sélection de l'échantillon. Avec l'échantillonnage non probabiliste, il faut faire des hypothèses



des UPF sont choisies par échantillonnage avec PPTF, l'application dans les UPF choisies de taux de sous-échantillonnage qui sont inversement proportionnelles aux tailles estimées des UPF produit un échantillon EPSEM global des visites. En général, un des attrait de l'échantillonnage avec PPTF (avec des estimations raisonnables de la taille) est que la taille des sous-échantillons des UPF ne varie pas beaucoup d'une UPF à l'autre. Cette caractéristique est particulièrement intéressante pour le travail sur le terrain lors d'enquête portant sur les personnes de passage. Quand des UPF portant sur le temps/l'endroit sont obtenues par échantillonnage avec PPTF, on ne peut appliquer l'échantillonnage de configuration pour une stratification en profondeur. On peut plutôt employer une sélection contrôlée à cette fin (Goodman et Kish 1950; Hess et coll. 1975).

Une considération importante dans tout plan d'échantillonnage à deux degrés est l'affectation de l'échantillon entre les unités primaires et les unités secondaires d'échantillonnage, c'est-à-dire, combien d'UPF choisir et combien d'éléments choisir dans chaque UPF sélectionnée. Dans le cas des enquêtes sur les personnes de passage, les procédures à utiliser sur le terrain ainsi que la nature du flux à l'intérieur des UPF ont un effet considérable sur le terrain affectation. Le but du plan d'échantillonnage est d'utiliser pleinement les travailleurs sur le terrain affectés à une UPF sélectionnée tout en maintenant un échantillon probabiliste des personnes qui entrent dans l'endroit (ou qui en sortent) pendant l'intervalle de temps échantillonné.

Dans de nombreuses enquêtes portant sur les personnes de passage on utilise des questionnaires à remplir soi-même; dans ce cas, le travail sur le terrain pour le plan d'échantillonnage à deux degrés décrit plus haut est composé du comptage des personnes au moment où elles entrent dans l'endroit échantillonné (ou lorsqu'elles en sortent) pendant l'intervalle de temps, du choix de chaque  $k^e$  personne pour un échantillon systématique et du fait de demander aux personnes sélectionnées de remplir le questionnaire. Si le flux n'est pas dense et s'il est réparti également pendant l'intervalle de temps, il se peut qu'un seul travailleur sur le terrain puisse remplir toutes les tâches. Quand c'est le cas, l'intervalle d'échantillonnage  $k$  peut être choisi de façon à donner à ce travailleur suffisamment de temps pour effectuer toutes les tâches sans travailler à la limite de ses possibilités. Si, toutefois, le flux est dense, que ce soit de façon constante ou intermittente, il se peut que l'on doive employer deux travailleurs, un qui n'aurait qu'à compter les entrants (ou les sortants) et à déterminer les personnes échantillonnées et le second pour remettre les questionnaires et dire aux répondants comment ils doivent le remplir et le retourner. Quand le travail sur le terrain est organisé de cette façon, on peut choisir l'intervalle d'échantillonnage pour que le second travailleur soit pleinement occupé, tout en s'assurant qu'il est en mesure de distribuer le questionnaire à toutes (ou presque toutes) les personnes choisies. La non-réponse peut être une préoccupation importante quand la collecte des données se fait à l'aide d'un questionnaire à remplir soi-même. Il est souvent possible de maintenir la non-réponse à un niveau acceptable quand les personnes choisies remplissent et remettent le questionnaire sur place. Toutefois, quand on leur remet le questionnaire en leur demandant de le remplir plus tard et de le retourner par la poste, le niveau de non-réponse peut être très élevé et, de plus, il n'existe généralement pas de façon d'effectuer le suivi des non-répondants.

Quand on a recours à des interviews comportant l'interrogation directe pour recueillir des données, l'équipe chargée du travail sur le terrain pour une UPF comprend généralement un compteur et un petit groupe d'intervieweurs. La taille de l'équipe d'intervieweurs dépend de la régularité du flux et de la durée de l'interview. Puisque les personnes de passage ne voudront vraisemblablement pas être retardées pendant une longue période, la majorité des interviews seront nécessairement courtes. Il est toutefois possible de réaliser des interviews plus longues si les personnes échantillonnées sont en attente, comme lorsqu'elles sont dans une file d'attente ou dans une salle d'attente d'un aéroport. Le choix de l'intervalle d'échantillonnage doit être tel qu'il y a toujours (ou presque toujours) un intervieweur libre pour interviewer la prochaine personne échantillonnée et que les intervieweurs ne passent pas trop de temps à attendre que

## 2. ÉCHANTILLONNAGE DANS LE TEMPS ET DANS L'ESPACE

Il sera utile de considérer un exemple particulier pour décrire le plan d'échantillonnage général dans le temps et dans l'espace servant à échantillonner des flux de populations humaines. Supposons qu'on doit réaliser une enquête auprès des visiteurs d'une exposition de sculptures tenue pendant l'été dans un parc municipal afin de trouver les caractéristiques socio-économiques des visiteurs, pour déterminer comment ils ont eu connaissance de l'exposition, quels moyens de transport ils ont utilisés pour se rendre au parc et, peut-être, quelles sont leurs opinions à propos de l'exposition. Supposons que l'exposition se déroule du 1<sup>er</sup> avril au 30 septembre de l'année en question, qu'elle est ouverte de 10 h à 18 h tous les jours et qu'il y a trois endroits où les visiteurs peuvent entrer sur le terrain de l'exposition et en sortir.

On considère généralement que la base de sondage employée pour une enquête de ce genre est une liste d'unités primaires d'échantillonnage (UPÉ) fondée sur l'intervalle de temps/l'endroit. On construit cette base de sondage en divisant la période de temps de l'enquête en un ensemble d'intervalles de temps pour chaque endroit. Une façon simple de construire les UPÉ pour l'exemple que nous étudions consisterait à diviser chaque jour d'exposition à chaque endroit en deux intervalles de temps, un allant de 10 h à 14 h et l'autre de 14 h à 18 h. Une façon plus complexe de construire les UPÉ pourrait comprendre l'utilisation d'intervalles de temps de longueur différente pour des jours différents et (ou) à des endroits différents. Une fois les UPÉ définies, on emploie souvent un plan d'échantillonnage à deux degrés. Dans la première étape, un échantillon d'UPÉ est choisi et, dans la deuxième étape, un échantillon de visiteurs est tiré, habituellement par échantillonnage systématique, dans les UPÉ échantillonnées.

Les spécifications réelles du plan d'échantillonnage utilisé pour une enquête visant les personnes de passage et qui utilise la base de sondage à deux degrés dépendent des caractéristiques de la population mobile à l'étude ainsi que des procédures utilisées pour faire la collecte des données d'enquête. Une caractéristique clé est la nature du flux de la population mobile. En particulier, y a-t-il une variabilité prévisible dans le taux du flux parmi les UPÉ? Par exemple, le flux à un endroit est-il plus élevé que celui que l'on retrouve à un autre endroit, ou les flux pour certains intervalles de temps (disons les samedis dans l'après-midi) sont-ils plus élevés que pour d'autres? De plus, le flux dans une UPÉ est-il continu pendant tout l'intervalle de temps ou est-il inégal, avec des visiteurs qui arrivent (ou qui partent) en groupes importants? Ces deux aspects du flux ont un effet sur le plan d'échantillonnage utilisé pour l'enquête.

Si le flux est assez uniforme parmi les UPÉ et si les intervalles de temps dans les UPÉ sont identiques, alors le nombre de visiteurs par UPÉ est approximativement constant. Dans ce cas, on peut échantillonner les UPÉ avec des probabilités égales et appliquer un taux de sous-échantillonnage constant dans les UPÉ choisies pour produire un échantillon avec probabilités égales, ou EPSEM, de visites. Les UPÉ peuvent être classées en deux dimensions ou plus (p. ex., le jour de la semaine, l'heure du jour et l'endroit) et on peut obtenir un échantillon bien équilibré entre ces dimensions à l'aide de l'échantillonnage de configuration (Yates 1981; Cochran 1977 et Jessen 1978).

Dans nombre de cas, le niveau du flux varie entre les UPÉ d'une façon qui est partiellement prévisible. Par exemple, on peut savoir que la fréquentation de l'exposition de sculptures est généralement plus élevée lors du dernier poste du travail de chaque jour et au cours des fins de semaine et qu'elle est particulièrement faible les lundis. Ainsi, les UPÉ comprennent un nombre différent de visiteurs, c'est-à-dire, qu'ils sont de taille différente. La procédure habituelle utilisée pour traiter des UPÉ de taille différente consiste à les échantillonner avec des probabilités proportionnelles à leur taille (PPT), ou avec des probabilités proportionnelles à leur taille estimée (PPTE). Dans le contexte actuel, la taille réelle des UPÉ n'est pas connue à l'avance, il faut donc utiliser des tailles estimées. L'échantillonnage des UPÉ avec PPT ou PPTE donne de bons résultats pourvu qu'on puisse faire des estimations raisonnables de leur taille. Quand



# L'Echantillonnage des flux de populations humaines mobiles

GRAHAM KALTON<sup>1</sup>

## RÉSUMÉ

On fait souvent des enquêtes sur des flux de populations de personnes comme celles qui vont dans les musées, les bibliothèques et les parcs; les électeurs; les personnes qui font des courses; les malades en consultation externe; les voyageurs venant de l'étranger et les passagers des automobiles. Les plans d'échantillonnage d'enquêtes de ce genre prévoient en général un échantillonnage dans le temps et dans l'espace. Dans cet article on passe en revue et on illustre les méthodes utilisées pour sonder les flux de populations humaines.

**MOTS CLÉS:** Populations mobiles; sondages des votants; enquêtes de circulation; échantillonnage dans le temps et dans l'espace; échantillonnage systématique.

## 1. INTRODUCTION

La majorité des enquêtes portant sur les populations humaines sont basées sur les ménages, on utilise habituellement un échantillon de ménages choisi au moyen d'un plan d'échantillonnage à plusieurs degrés puis on échantillonne des particuliers dans les ménages sélectionnés. L'enquête-ménage est une méthode puissante employée pour recueillir des données sur un grand nombre de caractéristiques relatives à la population, comme des caractéristiques sociales, démographiques et économiques ainsi qu'en matière de santé et les opinions et attitudes de la population. Toutefois, la méthode n'est pas aussi efficace pour étudier les caractéristiques de populations mobiles. On peut distinguer deux types de populations mobiles: les personnes qui n'ont pas de domicile fixe, comme les nomades et les sans-abri, ainsi que les gens en général qui font partie de la population mobile à l'étude parce qu'ils se trouvent momentanément à un endroit, comme les visiteurs qui vont dans des bibliothèques et des parcs, les électeurs dans les bureaux de scrutin, les personnes qui font leurs courses, les malades en consultation externe, les voyageurs et les passagers des automobiles. Dans le présent article, on passe en revue certaines questions portant sur les plans d'échantillonnage utilisés pour cette dernière catégorie de population mobile.

Bien que de nombreuses enquêtes portent sur les flux de populations humaines mobiles, les ouvrages généraux sur l'échantillonnage traitent peu des questions relatives à l'échantillonnage dans le cas des populations mobiles. L'objet du présent article est de décrire les plans d'échantillonnage généralement adoptés pour les enquêtes sur les flux de population humaine, de discuter de certains des problèmes particuliers qui se posent en matière d'échantillonnage et d'illustrer la gamme d'applications pour de telles enquêtes. Dans la section suivante de l'article nous passons en revue le plan général d'échantillonnage dans le temps et dans l'espace utilisé pour échantillonner des personnes mobiles et certaines des questions portant sur l'emploi de ce plan d'échantillonnage dans des situations particulières. La section 3 illustre ensuite l'application du plan d'échantillonnage dans une gamme de situations différentes. La section 4 renferme des conclusions.

<sup>1</sup> Graham Kalton, Survey Research Center, University of Michigan, Ann Arbor, Michigan 48106-1248, U.S.A.





On passe, par ailleurs, de  $\hat{p}^\circ$  à  $\hat{p}$  par une estimation du maximum de vraisemblance. Dans les conditions asymptotiques gaussiennes, celle-ci s'identifie à la minimisation de la forme quadratique  $(\hat{p}^\circ - \hat{p})' \Delta^{-1} (\hat{p}^\circ - \hat{p})$  sous les contraintes  $A\hat{p} = Ap$ . Comme  $\hat{p}^\circ$  varie dans le sous-espace affine  $L + V_0$  parallèle à  $L$ , que la minimisation en question revient à projeter  $\hat{p}^\circ$  sur  $L$  orthogonalement pour la métrique  $\Delta^{-1}$  c'est-à-dire le long de  $\Delta(L^\perp)$ , il suit qu'on a  $\hat{p} = \hat{p}^\circ - n^{*-1/2} V_0$  dans les conditions asymptotiques. Le vecteur aléatoire  $\hat{p}$  est donc translaté de  $\hat{p}^\circ$ , sans biais et de même matrice de covariance que  $\hat{p}^\circ$  et donc que  $n^{*-1/2} U$ . Finalement, on a:

$$E \left( \sum^c \hat{p}_c E_c \right)^2 = E (\hat{p}' \bar{E})^2 = \frac{1}{n^*} \sum^c p_c E_c^2$$

comme dans le cas précédent.

### BIBLIOGRAPHIE

CASSEL, C.M., SÄRNDALE, C.E., et WRETMAN, J.H. (1977). *Foundations of Inference in Survey Sampling*. New York: Wiley & Sons.

COCHRAN, W.G. (1977). *Sampling Techniques*. New York: Wiley & Sons.

DESABIE, J. (1965). *Théorie et pratique des sondages*. Paris: Dunod.

DEVILLE, J.C., et SÄRNDALE, C.E. (1990). Calibration estimators and generalized raking techniques. Manuscrit soumis pour publication.

GOURIEROUX, C. (1981). *Théorie des sondages*. Paris: Economica.

MADOW, W.G., OLKIN, I., et RUBIN, D.B., (éds.) (1983). *Incomplete Data in Sample Surveys*. New York: Academic Press.

RAO, J.N.K. (1976). Unbiased variance estimation for multistage designs. *Sankhyā*, Série C, 37, 133-139.

SMITH, T.M.F. (1983). On the validity of inferences from non-random samples. *Journal of the Royal Statistical Society, A*, 146, 394-403.

WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

avec  $A$  matrice à  $l = \sum^q l^q - (\bar{Q} - 1)$  lignes et  $k = \Pi^q l^q$  colonnes, composée de 1 et de 0 pour traduire les contraintes. Ceci exprime aussi le fait que  $U$  varie dans le noyau  $L$  de l'opérateur défini par la matrice  $A$ . La loi (asymptotique) de  $U$  est donc celle d'un vecteur gaussien  $W$  centré de matrice des covariances égales à  $\Delta$ , conditionnellement à  $AW = 0$ . On est ramené à évaluer la variance du produit scalaire  $U'\bar{E}$  où  $\bar{E}$  est le vecteur des  $E_c$ .

Remarquons les deux points suivants:

- Les contraintes sur les  $E_c$  données au résultat 1 se traduisent matriciellement par  $A\Delta\bar{E} = 0$ . Autrement dit  $\Delta\bar{E}$  est un vecteur de  $L = \text{Ker}A$ , ou encore  $\bar{E}$  est un vecteur de  $\text{Ker}(A\Delta)$ .
- Soit  $P$  le projecteur de  $\mathbb{R}^k$  sur  $L$  orthogonal dans la métrique  $\Delta^{-1}$ .  $P$  vérifie les relations:

$$\begin{aligned} & \bullet \text{ } x \in L, Px = x; \text{Im } P = L \\ & \bullet \text{ } Py = 0 \Leftrightarrow x \in L, x' \Delta^{-1} y = 0; \text{Ker } P = \Delta(L^\perp), \end{aligned}$$

où  $L^\perp$  est le supplémentaire orthogonal de  $L$  dans la métrique naturelle.

Les vecteurs gaussiens  $PW$  et  $(1 - P)W$  varient respectivement dans  $L$  et dans  $\Delta(L^\perp)$ , leur somme est égale à  $W$ . De plus ils sont indépendants; en effet leur matrice de covariance vaut  $E(PW)((1 - P)W)' = P\Delta(1 - P')$ . Or  $P'$  est le projecteur de noyau  $L^\perp$  et d'image  $\Delta(L^\perp)^\perp$ . L'image du projecteur  $(1 - P')$  est donc  $L^\perp$ , celle de  $\Delta(1 - P')$  est  $\Delta(L^\perp)$ , c'est-à-dire le noyau de  $P$ , c.q.f.d.

Il faut maintenant évaluer la variance de  $\sum^c n_c E_c = U'\bar{E}$ . Or, d'après ce qui précède, on peut écrire  $W = U + V$ , avec  $U$  et  $V$  indépendants. La loi de  $W$  conditionnellement à  $W \in L$  n'est autre que la loi de  $W$  conditionnellement à  $V = 0$ .

Par ailleurs on a:

$$V'\bar{E} = (\Delta^{-1}V)'(\Delta E).$$

Comme  $\Delta E$  est dans  $L$  et que  $V$  varie dans  $\Delta(L^\perp)$ , le produit scalaire ci-dessus est nul. On en déduit que:

$$\text{Var}(U'\bar{E}) = \text{Var}(W'\bar{E}) = \bar{E}'\Delta\bar{E} = \sum^c p_c E_c^2.$$

La variance asymptotique de  $N/n^* \sum^c n_c E_c$  vaut donc

$$\frac{N^2}{n} \sum^c p_c E_c^2 = \frac{n}{N} \sum^c N_c E_c^2.$$

#### 4. Cas des sondages par quotas "non-proportionnels"

Complétons la réduction asymptotique précédente. Maintenant le vecteur  $\hat{p}^\circ$  des  $n_c/n^*$  est contraint par

$$A\hat{p}^\circ = Ap + n^{*-1/2}AV_0,$$

où  $Ap$  est le vecteur (à  $l$  dimensions) des "quotas proportionnels" et  $V_0$  l'unique vecteur (à  $k$  dimensions) de  $\Delta(L^\perp)$  tel que  $A(p + n^{*-1/2}V_0)$  soit le vecteur des quotas imposés. De ce fait  $U_1 = n^{*1/2}(\hat{p}^\circ - p)$  peut-être, comme au paragraphe précédent, analysé comme un vecteur gaussien  $W = U + V$  conditionnel à  $V = V_0$ . Par suite  $EU_0 = V_0$  et la matrice des covariances de  $U_0$  est la même que celle de  $U$ .



Un sondage aléatoire simple sans remise (SASSR) de taille fixe  $n$  est un échantillonnage SR conditionnellement au fait que la taille totale de l'échantillonnage est  $n$ .  
 Un sondage aléatoire simple avec remise (SASAR) de taille fixe  $n$  est un échantillonnage AR conditionnellement au fait qu'on a fait  $n$  observations, c'est-à-dire que  $\sum_k v_k = n$ .  
 Dans le cas SR, la loi du vecteur  $n_c$  est donnée par:

$$\Pr(\{n_c\}) = \prod_{N_c}^{n_c} f^{n_c} (1 - f)^{N_c - n_c}.$$

Dans le cas AR, on a:

$$\Pr(\{n_c\}) = \prod_{N_c}^{n_c} \frac{(N_c f)^{n_c}}{n_c!} \exp(-fN_c).$$

Dans les deux cas les variables  $n_c$  sont indépendantes.

Dans le cas SR contraint par  $\sum n_c = n$ , la loi des  $n_c$  est une hypergéométrique:

$$\Pr(\{n_c\}) = \prod_{N_c}^{n_c} \binom{N_c}{n_c} \binom{n}{N}^{-1}.$$

Dans le cas AR contraint, c'est une multinomiale:

$$\Pr(\{n_c\}) = \prod_{N_c}^{n_c} p_c^{n_c}/n_c!.$$

Le modèle de plan de sondage retenu pour la méthode des quotas à la section 4 correspond à des contraintes sur ces deux schémas, ou, ce qui est équivalent, à des contraintes sur les schémas SR et AR.  
 Si on suppose que  $N$  tend vers l'infini, que  $f$  tend vers zéro, et que  $n^* = fN$  tend vers l'infini, alors, dans les deux schémas, la loi des  $n_c = n^* - \frac{1}{2} (n_c - fN_c) = n^*_{1/2} (p_c^2 - p_c)$ , avec  $p_c^2 = n_c/n^*$ , tend vers une loi normale multidimensionnelle avec les  $n_c$  indépendantes, d'espérances nulles et de variances égales aux  $p_c$ .

### 3. Cas des sondages "proportionnels"

On a alors  $N_c = N/n_c$  de sorte que la quantité dont on cherche la variance est:

$$\frac{N}{n_c} n^*_{1/2} \sum^c n_c E_c,$$

où le vecteur des  $n_c$  suit une loi normale centrée de matrice des covariances diagonale  $\Delta = \text{diag}(p_c)$ , contrainte par les relations qui expriment les quotas:

$$\sum^{q_c=i} n_c = 0 \text{ pour } q = 1 \text{ à } Q, \quad i = 1 \text{ à } I^q \text{ si } q = 1, \quad i = 1 \text{ à } I^q - 1 \text{ si } q = 2 \text{ à } Q.$$

Ces relations s'écrivent, en notant  $U$  le vecteur des  $n_c$ :

$$AU = 0,$$

ANNEXE

Démonstration des résultats en 4.4

1. Notations et résultats

Pour traiter la question de façon générale nous aurons besoin de certaines notations com-modes. On dispose de  $\bar{Q}$  variables qualitatives dont les modalités sont indexées de 1 à  $I_q$  pour  $q = 1$  à  $\bar{Q}$ . On note  $c$  une "cellule", c'est-à-dire une suite de  $\bar{Q}$  indices, le  $q^{\text{ième}}$  pouvant valoir de 1 à  $I_q$ , et  $q_c$  la valeur du  $q^{\text{ième}}$  indice ( $q^{\text{ième}}$  projection de  $c$ ); dans une population finie  $U$  d'effectif  $N$ ,  $U_c$  est la population des individus classés dans la cellule  $c$  d'effectif  $N_c$ . La quantité  $N_{i+q}^{\text{'}} = \sum_{q_c=i} N_c$  est la marge de la table de contingence  $\bar{Q}$ -dimensionnelle dont les cellules sont les  $c$ , pour la  $i^{\text{ième}}$  modalité de la  $q^{\text{ième}}$  variable. On pose

$$Y_c = \frac{1}{N_c} \sum_{k \in U_c} Y_k.$$

On a le:

**Résultat 1 :** La variable  $Y_k (k \in U)$  peut être paramétré par les nombres  $A_{q_c}^q, E_c$  et  $R_k$  par:

$$Y_k = Y_c + R_k \text{ si } k \in U_c. \text{ On a } \sum_{U_c} R_k = 0 \text{ pour tout } c.$$

$$Y_c = \sum_{\bar{Q}} A_{q_c}^q + E_c \text{ avec } A_{I_q}^q = 0 \text{ pour } q = 2 \text{ à } \bar{Q} \text{ et}$$

$$\sum_{q_c=i}^{q_c=\bar{Q}} N_c E_c = 0 \text{ pour } q = i \text{ à } \bar{Q} \text{ et } i = 1 \text{ à } I_q.$$

Ces nombres proviennent de la minimisation de:

$$\sum_{\bar{Q}} \left( Y_k - \sum_{\bar{Q}} A_{q_c}^q(k) \right)^2 = \sum_c N_c \left( Y_c - \sum_{\bar{Q}} A_{q_c}^q \right)^2.$$

Soit maintenant un échantillon  $s$ . On note avec des  $n$  les quantités analogues dans l'échan-tillon à ce qu'on a déjà indiqué dans la population avec des  $N$ .  
On suppose  $s$  tiré par sondage aléatoire simple (avec ou sans remise) selon un schéma à probabilités égales contraint par des marges  $n_{i+q}^{\text{'}}$  ( $q = 1$  à  $\bar{Q}$ ,  $i = 1$  à  $I_q$ ), les quotas. Le but de cette annexe est de montrer le:

**Résultat 2 :** La variance de  $\sum_c N_c E_c$  est approximativement égale à  $1/n \sum_c N_c E_c^2$  quand  $n$  et  $N/n$  deviennent arbitrairement grands.

La suite donnera une formulation précise à ce résultat.

2. Schéma d'échantillonnage et réduction asymptotique

Considérons les deux modèles d'échantillonnages SR et AR suivant:

SR: Échantillonnage bernouillien. Chaque unité parmi les  $N$  appartient à  $s$  avec probabilité  $f$ , les  $N$  tirages étant indépendants.  
AR: Chaque unité est tirée un nombre  $v_k$  de fois;  $v_k$  suit une loi de Poisson de paramètre  $f$ . Les  $v_k$  sont des variables indépendantes.

La statistique officielle, à l'opposé, est chargée d'élaborer des données utilisables par l'ensemble du corps social, susceptibles, en particulier, de servir d'éléments pour l'arbitrage de conflits entre divers groupes, divers partis, voire diverses classes sociales. Le recours à des modèles statistiques, économétriques en particulier, décrivant le comportement des agents économiques, peut se révéler assez dangereux, partiel, influencé par une théorie économique contestable, ou contestée. La statistique officielle ne doit tolérer aucun biais incontrôlable dans sa production. Elle se doit de réaliser des enquêtes par sondage par des méthodes probablistes. Il n'y a pas réellement opposition entre les enquêtes par quotas et les techniques ayant recours à un aléatoire contrôlé, mais, bien au contraire complètementarité. À preuve, les statistiques qui servent à construire les quotas sont elles-mêmes très souvent tirées de grosses enquêtes réalisées par les Instituts Nationaux de Statistique. Les techniciens des enquêtes par quotas admettraient mal que ces données soient élaborées autrement que par des méthodes probablistes confirmées et bien théorisées.

## REMERCIEMENTS

Je remercie bien sincèrement l'arbitre et le rédacteur associé du travail positif qu'ils ont réalisé et qui a contribué à l'amélioration de cet article.



Une technique de raking ratio permet donc de calculer des estimations  $f_c^*$  et  $w_i^*$  des  $r_c$  et  $w_i$ . On en déduit des estimateurs  $N_i^c = n_c^* w_i^* f_c^{-1}$  des effectifs du croisement  $(i, c)$ . On en déduit aussi un estimateur du total de  $Y$ :

$$Y^{NR} = \sum_{ic} N_i^c y_i^c = \sum_{ic} r_c^{-1} w_i n_i^c y_i^c,$$

où  $y_i^c$  est la moyenne des  $Y_c$  de l'échantillon classés dans la catégorie  $(i, c)$ . Ainsi, les techniques d'estimation par calage devraient permettre un traitement honorable de la non-réponse y compris dans des enquêtes par quotas.

### 5.2 Quelques points de comparaison avec des sondages probabilistes

La méthode des quotas, quelque soit la façon dont on essaie de l'envisager, réclame la formation d'un modèle hypothétique qu'on plaque sur les données. À l'inverse, un sondage probabiliste ne dépend d'aucun modèle, en principe. En pratique, l'échantillonnage d'un sondage probabiliste est un modèle auquel la réalité de la collecte des données essaie de se conformer. On sait bien, en effet, que dans tout sondage probabiliste, quelques accommodements de détail doivent être pris avec le modèle (exclusion d'office de certaines unités, remplacement de certaines après tirage mais avant collecte, etc.). On peut, cependant, affirmer sans risque que les biais statistiques sont toujours beaucoup plus faibles dans les tirages probabilistes qu'avec la méthode des quotas. En revanche, les quotas permettent d'utiliser au stade de l'échantillonnage une information auxiliaire qui n'est pas mobilisable dans un tirage probabiliste. Il en résulte que la variance d'un échantillonnage par quotas, est du genre de celle d'une estimation par régression et qu'elle est donc plus faible en règle générale que celle qui résulte d'un sondage probabiliste associé à son estimation de valeurs diluées standard. Biais dû au modèle associé à une faible variance, contre absence de biais, tel est le bilan. On peut tirer de cette approche deux types de conclusions:

**5.2.1** La précision dépend avant tout de la taille des échantillons. Dans le cas de faibles échantillons, le sondage probabiliste va donner de piètres résultats en moyenne et le biais d'un sondage par quotas sera plus tolérable que l'imprécision du sondage probabiliste. Pour de gros échantillons au contraire, la méthode des quotas aura un biais évident incompatible avec l'intervalle de confiance sans biais du sondage probabiliste.

On fixe la limite entre les deux méthodes? La théorie peut difficilement être affirmative. En revanche, la pratique des instituts français propose une solution à cette question: la plupart des enquêtes nationales par quotas sont réalisées sur des échantillons de 1,000 à 2,000 individus. En revanche, aucune enquête probabiliste nationale ne mobilisera moins de 5,000 unités. Il paraît légitime de dire qu'un effectif de 2,500 à 3,000 enquêtes fixe une limite pratique entre les deux familles de méthodes.

### 5.2.2 Statistique officielle ou marketing

Tout modèle spéculatif constitue, dans une enquête, une prise de risque méthodologique. Ce risque peut être parfaitement légitime si les utilisateurs en sont conscients, s'ils ont ratifié la spéculation qui a conduit à la spécification d'un modèle. C'est typiquement ce qui se produit, au moins de façon implicite, dans les enquêtes de marketing: un organisme, société, administrateur ou association, passe commande d'une enquête par sondage avec une société d'études. Un contrat marque l'accord entre les deux parties sur la réalisation de l'enquête, son prix, les délais de livraison des résultats et la **méthodologie employée**. Dans cette méthodologie il y a les modèles utilisés pour formaliser l'échantillonnage ou le comportement de la population. La méthode des quotas peut donc être, de ce point de vue, tout à fait légitime.

4.5.2 Cas d'un sondage à deux degrés

Supposons un sondage à deux degrés (éventuellement à l'intérieur d'une strate où les effectifs des variables de quotas sont connus). Si les effectifs des variables de quotas sont connus au niveau de chaque unité primaire il n'y a pas de problème. La théorie en 4.4 permet de former un estimateur du total de  $X$  dans chaque unité primaire, ainsi que de calculer sa variance et un estimateur de celle-ci. Ces quantités peuvent donc être utilisées pour former un estimateur de  $X$  ainsi qu'un estimateur de précision (cf Rao 1975).

Si les effectifs des critères de quotas sont inconnus au niveau des UP mais connus seulement au niveau de la strate, on a de nouveau un problème de correction impossible. Toutefois, le mal doit généralement être limité si les UP sont relativement semblables entre elles: la structure de chaque UP est proche de celle de la strate toute entière et les corrections à faire pour chaque UP sont voisines de celles qu'on doit mettre en oeuvre au niveau de la strate.

4.5.3 En conclusion

Dans le cas d'un sondage complexe stratifié à plusieurs degrés, la méthode des quotas peut être utilisée comme ultime méthode d'échantillonnage si la stratification a été réalisée de façon efficace en regroupant des unités primaires assez semblables entre elles et si on applique dans chaque UP des quotas dérivés de données relatives à sa strate.

Dans la mesure où l'hypothèse d'un échantillonnage aléatoire simple contraint dans chaque UP peut sembler assez satisfaisante, la méthode des quotas reçoit une justification indépendante de tout modèle de superpopulation.

5. CONCLUSIONS ET PROBLEMES OUVERTS

5.1 Comment prendre en compte la non-réponse?

Comme nous l'avons déjà signalé, cette question est la limitation la plus importante de notre théorie. Quand on échantillonne par la méthode des quotas, on n'a, en principe, aucune information sur la population qui refuse de répondre à l'enquête et on se trouve démuní de l'information individuelle au sujet des non-répondants. La situation n'est, cependant, peut être pas si désespérée qu'on pourrait le croire. Illustrons cette intuition par un exemple très simplifié.

On a réalisé une enquête par quotas simples chargeant l'échantillon de  $n_i$  individus de la catégorie  $i$  d'effectif  $N_i$ . Un modèle (admis) de non-réponse postule une probabilité  $r_c$  de réponse si un individu appartient à une catégorie  $c$  d'effectif  $N_c$ . L'effectif (inconnu) du croisement entre la catégorie  $i$  de quota et la classe  $c$  du modèle de non-réponse est noté  $N_{ic}^*$ . L'effectif susceptible de répondre dans la catégorie  $i$  vaut donc  $N_{it} = \sum_c N_{ic}^*$ . En fixant un quota  $n_i$  dans cette catégorie, dans le cadre du modèle (4.1), on obtient une probabilité d'inclusion dans l'échantillon égale à  $w_i^{-1} = n_i/N_{it}$ . Dans l'échantillon, on recueille  $n_i^*$  individus appartenant au croisement ( $i, c$ ) des deux catégorisations. Cette quantité est aléatoire et son espérance vaut  $N_{ic}^* r_c w_i^{-1}$ . Si on cherche à estimer les  $N_{ic}^*$ , on résoudre les équations estimantes déduites des relations:

$$N_i^* = n_i^* w_i r_c^{-1},$$
$$\sum_i N_i^* = N_i,$$
$$\sum_i N_i^* = N_c.$$

L'estimateur de la variance est alors:

(4.4.3)

$$\widehat{\text{Var}}(Y_0) = (N^2/n) \sum_{ij} p_{ij} (E_{ij}^2 + (a_i + b_j)^{-1} s_{ij}^2).$$

Dans le cas des quotas proportionnels aux effectifs cette expression se simplifie en:

(4.4.4)

$$(N^2/n) \sum_{ij} n_{ij} (E_{ij}^2 + s_{ij}^2)/n.$$

Si les  $n_{ij}$  sont tous suffisamment grands pour qu'on puisse prendre  $n_{ij}/(n_{ij} - 1) = 1$  on voit que la somme de la formule n'est autre que la somme des carrés des résidus estimés dans l'ajustement par les MCO du modèle  $Y_k = A_i + B_j + \text{résidu}$ . La procédure d'estimation est alors simple:

- ajuster par les MCO le modèle additif sur données individuelles
- créer la variable  $e_k$  des résidus estimés
- $\widehat{\text{Var}}(Y_0) = (N^2/n) \cdot (1/n) \sum_s e_k^2$ .

Cette formule est exactement celle qui avait été proposée à la section 3 à partir du modèle de superpopulation, situation assez sympathique!

#### 4.4.3 Discussion des résultats

La variance se décompose en deux parts, l'une vue comme l'espérance du carré du biais conditionnel, la seconde comme l'espérance de la variance conditionnelle. Le premier terme ne dépend pas des quotas imposés à l'échantillon, mais seulement de la qualité de l'ajustement d'un modèle additif sur la variable d'intérêt. On diminue cette partie de la variance en choisissant des critères de quotas qui expliquent au mieux ce qu'on veut mesurer.

Le second terme, en revanche, dépend de la variabilité restante ( $N_{ij}^2 V_{ij}/n_{ij}$ ) et du nombre d'observations recueillies dans chaque cellule. La taille de l'échantillon étant fixe on doit donc chercher à rendre les  $n_{ij}$  les plus proches possible de l'allocation de Neyman:  $n_{ij} \propto N_{ij} V_{ij}^{1/2}$ . Ceci peut s'obtenir approximativement en surchargeant les quotas  $n_{i+}$  et  $n_{+j}$  qui correspondent à de grandes valeurs de  $V_{ij}$ . On peut ainsi, dans certains cas, améliorer sensiblement la précision des enquêtes par quotas.

### 4.5 Combinaison de la méthode des quotas avec les échantillonnages stratifiés où à plusieurs degrés

#### 4.5.1 Cas d'un sondage stratifié avec quotas dans chaque strate

Si les effectifs des critères servant à fabriquer les quotas sont connus dans chaque strate, la méthode qui vient d'être décrite permet de construire un estimateur sans biais sous l'hypothèse que l'échantillonnage fonctionne comme un SAS contraint dans chaque strate. Si l'allocation des quotas est proportionnelle aux effectifs de chaque strate, l'estimateur est l'estimateur naturel du sondage stratifié. Si on applique des quotas "nationaux" à chaque strate, une correction doit être faite par pondération.

En revanche, si les effectifs des variables de quotas sont inconnus au niveau des strates, on ne dispose d'aucun moyen de corriger les estimateurs des "effets de structures" relatifs à la stratification. Comme, de plus, la stratification a pour but de construire des quotas n'est alors dissemblables, ces corrections seraient généralement fortes. La méthode des quotas n'est alors pas à recommander (sauf, cf section 3, si la validité d'un modèle additif s'impose d'elle-même).



Conditionnellement aux  $n_{ij}$ , les  $N_{ij}$  sont constants et les sous-échantillons  $s_{ij}$  des sondages aléatoires simples indépendants. On a donc :

$$\text{Biais cond}(Y_{\bar{Q}}) = \sum_{ij} N_{ij} E_{ij} = N \sum_{ij} \hat{p}_{ij} E_{ij}$$

$$\text{Var cond}(Y_{\bar{Q}}) = \sum_{ij} N_{ij}^2 V_{ij}/n_{ij} \quad \text{ou} \quad V_{ij} = (1/N_{ij}) \sum_{k=1}^{U_{ij}} R_k^2.$$

Or (démonstration en annexe) on a le :

**Résultat 1 :**

$$\text{Var} \left( \sum_{ij} \hat{p}_{ij} E_{ij} \right) = 1/n \sum_{ij} p_{ij} E_{ij}^2.$$

Par ailleurs, l'espérance de  $\hat{p}_{ij}(a_i^* + b_j^*)^{-1}$  vaut, (à des termes en  $1/n$  près)  $p_{ij}(a_i^* + b_j^*)^{-1}$  où  $a_i^*$  et  $b_j^*$  sont solutions des équations (4.3.4) dans lesquelles on remplacerait les  $\hat{p}_{ij}$  par les  $p_{ij}$  exacts.

D'où le résultat :

**Résultat 2 :** La variance de l'estimateur des quotas  $Y_{\bar{Q}}$  est donnée par :

$$\text{Var}(Y_{\bar{Q}}) = (N^2/n) \sum_{ij} p_{ij} (E_{ij}^2 + (a_i^* + b_j^*)^{-1} V_{ij}).$$

Si les quotas sont proportionnels aux effectifs dans la population, on aura :

$$\text{Var}(Y_{\bar{Q}}) = (N^2/n) \sum_{ij} p_{ij} (E_{ij}^2 + V_{ij}).$$

#### 4.4.2 Estimation de la variance

La variance conditionnelle de  $Y_{\bar{Q}}$  s'estime immédiatement par :

$$\sum_{ij} N_{ij}^2 s_{ij}^2/n_{ij} = (N^2/n) \sum_{ij} \hat{p}_{ij} (a_i + b_j)^{-1} s_{ij}^2.$$

où  $s_{ij}^2$  est l'estimateur sans biais habituel de  $V_{ij}$ . L'espérance du carré du biais conditionnel vaut  $(N^2/n) \sum_{ij} p_{ij} E_{ij}^2$  et s'estime par  $(N^2/n) \sum_{ij} \hat{p}_{ij} E_{ij}^2$  où  $E_{ij} = y_{ij} - \hat{A}_i - \hat{B}_j$  avec  $\hat{A}_i$  et  $\hat{B}_j$  solutions de :

$$\sum_{ij} \hat{p}_{ij} (\hat{A}_i + \hat{B}_j) = \sum_{ij} \hat{p}_{ij} y_{ij} \quad (i = 1 \text{ à } I),$$

$$\sum_{ij} \hat{p}_{ij} (\hat{A}_i + \hat{B}_j) = \sum_{ij} \hat{p}_{ij} y_{ij} \quad (j = 1 \text{ à } J - I) \quad \text{avec} \quad B_J = 0.$$

(4.4.2)

Autrement dit on obtient l'estimation des  $E_{ij}$  en ajustant aux données un modèle ANOVA additif sans interaction, le critère d'ajustement étant celui des moindres carrés pondérés par les poids  $(a_i + b_j)^{-1}$ .

Les estimateurs des  $p_{ij}^0$  sont alors les  $\hat{p}_{ij}^0(a_i + b_j)^{-1}$  et l'estimateur cherché s'écrit:

$$\bar{Y}_{\hat{Q}} = (N/n) \sum_{ij} n_{ij}(a_i + b_j)^{-1} \hat{y}_{ij} = (N/n) \sum_s w_k X_k. \tag{4.3.5}$$

où  $w_k = (a_i + b_j)^{-1}$  est la pondération à appliquer à  $X_k$  dans le cas où  $k$  appartient à la cellule  $(i, j)$ . Cet estimateur est asymptotiquement sans biais sous le modèle de S.A.S. dans  $U$  comme le sont les estimateurs du maximum de vraisemblance. Les quotas  $n$  interviennent pas de façon explicite dans 3.3.4 mais ils influent sur les valeurs des  $a_i$  et  $b_j$ . Dans le cas habituel où les quotas marginaux sont "proportionnels" avec une fraction de sondage fixe  $f$ , la solution des équations 4.3.4 est évidente:  $a_i = 1$  pour tout  $i$  et  $b_j = 0$  pour tout  $j$ . L'estimateur du total vaut  $N\bar{y}$ , comme on pouvait s'y attendre et à la même expression que pour un sondage probabiliste à probabilités égales.

**Remarque.** L'utilisation du maximum de vraisemblance pour estimer les proportions est assez arbitraire. Un critère du type chi-2 (minimiser  $\sum_{ij} (p_{ij} - \hat{p}_{ij}^0)^2 / \hat{p}_{ij}^0$ ) rendrait linéaire le système (4.3.4).

#### 4.4 La variance de l'estimateur et son estimation

**4.4.1** Pour établir une formule de variance nous utiliserons la paramétrisation de la variable  $Y$  utilisée dans Deville et Särndal (1990) que nous énonçons sous forme d'un:

**Lemme:** Pour toute variable  $Y = (Y_k; k \in U)$  on peut choisir une paramétrisation définie de façon unique

$$Y_k = Y_{ij} + R_k \quad \text{si } k \text{ est dans la cellule } (i, j) \quad (k \in U_{ij}) \quad \text{avec} \quad \sum_{k \in U_{ij}} R_k = 0,$$

$$Y_{ij} = A_i + B_j + E_{ij} \quad \text{avec} \quad B_j = 0$$

$$\sum_{ij} N_{ij} E_{ij} = 0 \quad i = 1 \text{ à } I$$

$$\sum_{ij} N_{ij} E_{ij} = 0 \quad j = 1 \text{ à } J - I.$$

De fait les  $A_i$  et  $B_j$  sont les nombres qui minimisent la quantité  $\sum_U (Y_k - A_i - B_j)^2$  où, de façon équivalente,  $\sum_{ij} N_{ij} (Y_{ij} - A_i - B_j)^2$ .

On peut alors écrire:

$$\bar{Y}_{\hat{Q}} = (N/n) \sum_{ij} n_{ij} (a_i + b_j)^{-1} (A_i + B_j + E_{ij} + R_{ij}) \quad \text{où} \quad \bar{R}_{ij} = \sum_{s_{ij}} R_k / n_{ij}.$$

Compte tenu des équations 4.3.4 et du lemme on en déduit:

$$\bar{Y}_{\hat{Q}} - Y = \sum_{ij} N_{ij} (E_{ij} + \bar{R}_{ij}) \quad \text{avec} \quad \bar{N}_{ij} = (N/n) n_{ij} (a_i + b_j)^{-1}, \tag{4.4.1}$$

qui est l'expression de base pour le calcul de variance.

4.2 Quotas par cellule

Ce modèle d'échantillonnage se ramène à la stratification *a priori*. Son avantage pratique est de ne pas nécessiter une base de sondage où sont présentes les variables de stratification. Il est mis en oeuvre rigoureusement dans certains cas, par exemple celui d'un sondage par téléphone où on part d'une liste aléatoire de numéros non informatifs et où on réalise des enquêtes jusqu'à ce que les quotas soient satisfaits.

Les formules donnant estimations de précision sont donc celles qu'on trouve dans tous les manuels. Elles présentent une analogie certaine avec celles en 3.1 (voir Gouriéroux 1981).

4.3 Cas des quotas marginaux: généralités-estimateurs

Le modèle d'échantillonnage est celui du sondage aléatoire simple contraint par les quotas marginaux. Le S.A.S fournit des échantillons comportant des effectifs  $n_{ij}$  dans les différentes cellules qu'on peut voir comme un vecteur aléatoire (à valeurs entières) dans  $R_{IJ}$ . La contrainte des quotas signifie qu'on se limite au vecteur aléatoire conditionné par:

$$\sum_{j=1}^J n_{ij} = n_{i+} \quad (i = 1 \text{ à } I) \quad \text{et} \quad \sum_{i=1}^I n_{ij} = n_{+j} \quad (j = 1 \text{ à } J - 1),$$

c'est-à-dire variant dans un sous espace de dimension  $IJ - I - J + 1$ . Nous nous plaçons dans le cas où le taux de sondage global est négligeable et où la loi des  $n_{ij}$  peut être assimilée à une loi multinomiale ( $n, p_{ij} = N_{ij}/N$ ).

Conditionnellement aux  $n_{ij}$ , les  $y_{ij}$  estiment sans biais les  $Y_{ij}$ . L'idée est maintenant de construire un estimateur du total de  $Y$  en pondérant les  $y_{ij}$  par des estimateurs des  $N_{ij}$ , c'est-à-dire des  $p_{ij}$ . Si on choisit de maximiser la vraisemblance, celle-ci est proportionnelle à:

(4.3.1) 
$$\prod_{ij} p_{ij}^{n_{ij}}.$$

On doit donc maximiser

(4.3.2) 
$$\sum_{ij} n_{ij} \text{Log} p_{ij}$$

sous les contraintes

(4.3.3) 
$$\sum_{j=1}^J p_{ij} = p_{i+} \quad (i = 1 \text{ à } I) \quad \text{et} \quad \sum_{i=1}^I p_{ij} = p_{+j} \quad (j = 1 \text{ à } J - 1)$$

ce qui amène à résoudre le système en  $a_i, b_j$  (les  $p_{i+} = N_{i+}/N$  et  $p_{+j} = N_{+j}/N$  sont connus):

(4.3.4) 
$$\sum_{j=1}^J p_{ij}^{\circ} (a_i + b_j)^{-1} = p_{i+} \quad (i = 1 \text{ à } I) \\ \sum_{i=1}^I p_{ij}^{\circ} (a_i + b_j)^{-1} = p_{+j} \quad (j = 1 \text{ à } J - 1; b_J = 0),$$

avec  $p_{ij}^{\circ} = n_{ij}/n$  fréquence observée sur l'échantillon.



D'autre part, si de "bonnes" précautions d'échantillonnage sont prises,  $Nn_{ij}/n - N_{ij}$  devrait être voisin de 0 assez souvent.

Il est clair, en tout cas, que mieux le modèle additif "colle" ( $\gamma_{ij}$  petits) et plus le plan de sondage se rapproche de l'aléatoire, plus le biais a des chances de se réduire.

**3.3.2** Une autre façon d'envisager la fausseté du modèle, déjà signalée, est de ne plus admettre l'indépendance entre l'aléa d'échantillonnage et l'aléa du modèle additif. Ceci revient à dire que des modèles distincts doivent être développés pour des vecteurs ( $Y_k, k \in s$ ) et ( $Y_i, i \in s$ ). Cette façon de voir les choses a été souvent employée dans la littérature économétrique à laquelle nous renvoyons le lecteur. Il est clair que la prise de risque vis à vis des données devient alors énorme et, souvent, incompatible avec un travail objectif de statisticien.

**3.4 Quotas marginaux à taux inégaux**

Dans le cas de quotas par cellules on peut fixer arbitrairement les quotas de chaque cellule. Jusqu'ici, dans le cas de quotas marginaux, nous n'avons envisagé que le cas où les quotas étaient proportionnels aux effectifs de la population.

Dans de nombreux cas, toutefois, on est tenté de surreprésenter certaines catégories. Si on désire étudier, par exemple, les patrimoines des ménages, on désira fixer des quotas plus importants pour les ménages âgés d'une part (quotas par groupes d'âge) et pour ceux dont le chef est travailleur indépendant (quotas par catégories sociales).

Formellement, on impose donc à l'échantillon de respecter des effectifs  $n_{i+}$  et  $n_{+j}$  *a priori* quelconques (avec toutefois la somme des  $n_{i+}$  égale à la somme des  $n_{+j}$ ).

Dans ce cas, en utilisant toujours les MCO comme technique d'estimation, on trouve facilement que l'estimateur par prédiction du total est:

$$X_{\text{Pred}} = \sum_i N_{i+} \hat{\alpha}_i + \sum_j N_{+j} \hat{\beta}_j, \tag{3.4.1}$$

les  $\hat{\alpha}_i$  et  $\hat{\beta}_j$  vérifiant toujours les équations estimantes (3.2.2). Il est facile de voir que cet estimateur peut se mettre sous la forme:

$$X_{\text{Pred}} = \sum_{ij} (w_{i(1)}^i + w_{i(2)}^j) n_{ij} \hat{y}_{ij} = \sum_{ij} N_{ij} \hat{y}_{ij}.$$

Les quantités  $(w_{i(1)}^i + w_{i(2)}^j) n_{ij}$  apparaissent donc comme des estimations des effectifs des cellules ( $i, j$ ), idée qui sera largement exploitée dans la suite.

En revanche, la variance sous modèle de cet estimateur dépend de l'ensemble des  $n_{ij}$ , comme le montre un calcul un peu laborieux. La justification de la méthode des quotas évoquée précédemment ne fonctionne plus.

**4. MODÈLES POUR LE PLAN DE SONDAGE**

**4.1 Un modèle de plan de sondage**

L'idée est celle d'un sondage aléatoire simple (S.A.S) contraint par les quotas imposés. L'algorithme de tirage, tout à fait utopique, serait de tirer une suite d'échantillons aléatoires simples jusqu'à ce qu'on en rencontre un qui vérifie les quotas. Ainsi, chaque échantillon vérifiant les quotas à la même probabilité positive d'être tiré, les échantillons ne vérifiant pas les quotas ayant une probabilité nulle.

Cette vue de l'esprit cherche à modéliser le fait qu'un enquêteur va suivre correctement les consignes de dispersion des unités sondées données par son encadrement.

Mais

$$\sum_{ij} N_{ij}(\sigma_i^2 + \tau_j^2) = \sum_i N_{i+} \sigma_i^2 + \sum_j N_{+j} \tau_j^2$$

$$= (N/n) \left( \sum_i n_{i+} \sigma_i^2 + \sum_j n_{+j} \tau_j^2 \right)$$

d'où :

$$E(N\bar{y} - Y)^2 = (N^2/n) (1 - f) n^{-1} \left( \sum_{ij} n_{ij} (\sigma_i^2 + \tau_j^2) \right)$$

$$= (N^2/n) (1 - f) \left( \sum_i p_{i+} \sigma_i^2 + \sum_j p_{+j} \tau_j^2 \right)$$

$$\text{avec } p_{i+} = N_{i+}/N \text{ et } p_{+j} = N_{+j}/N.$$

L'estimation de la précision de  $E(N\bar{y} - Y)^2$  en découle. En effet  $s_{\bar{y}}^2$  a pour espérance sous modèle  $\sigma_i^2 + \tau_j^2$ . On obtient donc un estimateur sans biais de la précision par

$$\sum_{ij} n_{ij} s_{ij}^2 (N/n)^2 (1 - f)$$

si tous les  $n_{ij}$  sont supérieurs ou égaux à 2.

Cet estimateur est, formellement, identique à celui qu'on utiliserait dans une poststratification complète sur les cellules  $(i,j)$ . On peut aussi utiliser  $(N/n)^2 (1 - f) \sum_s e_k^2$  où les  $e_k$  sont les résidus estimés du modèle.

### 3.3 Et si le modèle est faux?

**3.3.1** Une première façon de voir la question est de plonger le modèle (3.2.1) dans le modèle général où la moyenne de  $Y_k$  dépend du couple  $(i,j)$ . On peut écrire cela sous la forme :

$$(3.3.1.1) \quad Y_k = \alpha_i + \beta_j + \gamma_{ij} + \epsilon_k,$$

avec les hypothèses habituelles sur les termes d'interaction  $\gamma_{ij}$  qui vérifient des contraintes d'identifiabilité :

$$(3.3.1.2) \quad \sum_j N_{ij} \gamma_{ij} = 0 \text{ et } \sum_i N_{ij} \gamma_{ij} = 0.$$

On a alors, de façon immédiate :

$$(3.3.1.3) \quad E(N\bar{y} - Y) = \sum_{ij} (N n_{ij}/n - N_{ij}) \gamma_{ij},$$

de sorte que l'estimateur est biaisé pour le modèle sauf si  $n_{ij} = f N_{ij}$  ce qui n'a aucune raison d'être réalisé. Ceci dit les termes de la somme (3.3.1.3) peuvent très bien se compenser, avec un peu de chance, car leurs signes sont *a priori* indéterminés.

3.2 Quotas marginaux - Cas "représentatif"

Dans ce paragraphe et dans toute la suite, nous nous bornerons au cas de quotas croisant 2 critères  $i$  et  $j$ . La généralisation à plus de 2 critères ne pose aucun problème particulier mais génère des notations très lourdes auxquelles on a préféré renoncer (voir l'annexe).

La situation est donc la suivante: les effectifs  $N_{i+}$  et  $N_{+j}$  des deux ventilations de l'univers sont connues. L'échantillonnage n'autorise que des échantillons de taille fixe  $n = fN$  comportant pour chaque  $i$   $n_{i+} = fN_{i+}$  individus et pour chaque  $j$   $n_{+j} = fN_{+j}$  individus.

Nous postulons dans la population, un modèle de type analyse de la variance formulé de la façon suivante:

Si  $k$  appartient à la cellule  $(i,j)$ :

(3.2.1) 
$$X_k = \alpha_i + \beta_j + \epsilon_k.$$

Les  $\epsilon_k$  sont centrées, indépendantes et on a  $\text{Var } \epsilon_k = \sigma_i^2 + \gamma_j^2$ .

Pour des raisons d'identification du modèle on pose  $\beta_j = 0$ .

Ceci équivaut à poser  $X_k = (\alpha_i + u_{ik}) + (\beta_j + v_{jk})$  où les  $u_{ik}$  et les  $v_{jk}$  sont indépendants et de variance respectives  $\sigma_i^2$  et  $\gamma_j^2$ .

On estime les  $\alpha_i$  et  $\beta_j$  par la méthode des moindres carrés ordinaires (MCO) car on ignore les valeurs des éléments de la variance; les  $\hat{\alpha}_i$  et  $\hat{\beta}_j$  sont solutions du système:

(3.2.2) 
$$\sum_{j=1}^J n_{ij} y_{ij} = n_{i+} \hat{\alpha}_i + \sum_{j=1}^J n_{ij} \hat{\beta}_j \quad (i = 1 \text{ à } I)$$
$$\sum_{i=1}^I n_{ij} y_{ij} = n_{+j} \hat{\beta}_j + \sum_{i=1}^I n_{ij} \hat{\alpha}_i \quad (j = 1 \text{ à } J - 1),$$

avec  $y_{ij}$  moyenne des  $X_k$  sur  $s_{ij}$  partie de l'échantillon dans la cellule  $(i,j)$ . L'estimateur par prédiction s'écrit alors:

$$X^{\text{Pred}} = \sum_{ij} (N_{ij} - n_{ij}) (\hat{\alpha}_i + \hat{\beta}_j) + \sum_{ij} n_{ij} y_{ij}.$$

**Résultat 1:** Sous le modèle (3.2.1), l'estimateur par prédiction utilisant les MCO est  $N_Y$ . On vérifie qu'il est sans biais pour le modèle c'est à dire que  $E(N_Y - Y) = 0$ .

**Preuve:** immédiate à partir de (3.2.2) et du fait que les quotas sont proportionnels aux effectifs dans la population.

**Résultat 2:** On a:

$$E(N_Y - Y)^2 = (N^2/n) (1 - f) n^{-1} \left( \sum_i n_{i+} \sigma_i^2 + \sum_{+j} n_{+j} \gamma_j^2 \right).$$

Cette quantité ne dépend pas de l'échantillon (puisque qu'elle ne dépend que des quotas). On a donc là, dans une certaine mesure, une justification de la méthode des quotas marginaux.

**Preuve:** Avec  $m_k = E X_k$  on a, en utilisant le caractère non biaisé de l'estimateur:

$$E(N_Y - Y)^2 = E \left( (N/n) \sum_s \epsilon_k - \sum_z \epsilon_l \right)^2 = (N/n)^2 \sum_{ij} n_{ij} (\sigma_i^2 + \gamma_j^2) - 2(N/n) \sum_{ij} n_{ij} (\sigma_i^2 + \gamma_j^2) + \sum_{ij} N_{ij} (\sigma_i^2 + \gamma_j^2).$$



## 2.4 Remarques sur les deux optiques appliquées à la méthode des quotas

a) Dans les deux cas, l'estimation sera efficace si la variable d'intérêt est bien expliquée par les indicatrices des catégories sur lesquelles on fonde les quotas, grosso modo, parce que les résidus d'ajustement de la régression seront petits.

b) Dans une enquête par quotas le "plan de sondage" est inconnu du statisticien. Celui-ci ne peut donc faire d'inférence sans recourir à un modèle. Ce peut être un modèle de comportement de la population (optique "modèle") qui l'oblige à prendre des responsabilités vis-à-vis de la nature de ce qu'il observe. Ce point de vue sera développé dans la troisième partie de ce papier. Ce peut être aussi une modélisation du plan de sondage, ce qui veut dire une prise de responsabilité vis-à-vis du fonctionnement du processus de collecte. Ce point de vue sera développé dans la quatrième partie de l'article.

Dans tous les cas la spéculation modélisatrice doit être mobilisée pour valider une forme d'inférence. La question est de savoir s'il est plus facile et plausible de modéliser le comportement des individus qu'on sonde que de modéliser le processus de recueil de l'échantillon (y compris dans ses aspects de contact entre enquêteur et enquêté).

c) À cet égard l'hypothèse faite en 2.3 d'indépendance entre des aléas dans la population et des aléas dans le processus de collecte est **cruciale**. Si l'échantillonnage est contrôlé par les statisticiens, cette hypothèse est garantie, sauf effet des non-réponses. Dans le cas de la méthode des quotas on n'a aucune garantie. Supposons par exemple que l'on désire mesurer des revenus  $X_k$ : la probabilité  $\pi_k$  de trouver  $k$  dans l'échantillon peut être très diminuée si  $X_k$  est grand. Autrement dit l'appartenance à l'échantillon (variable qui vaut 1 si  $k$  est dans  $s$  et 0 sinon) et le résidu du modèle de superpopulation  $\epsilon_k$  sont corrélés négativement. Cet exemple illustre bien le principal danger de la méthode des quotas, la théorie qui suit n'en tient pas compte.

## 3. THÉORIE DES QUOTAS AVEC MODÈLE DE SUPERPOPULATION

### 3.1 Quotas par cellule

On a une seule catégorisation en cellules  $i = 1$  à  $I$  d'effectifs connus  $N_i$ . Le modèle qu'on peut imaginer est le suivant:

$$(3.1.1) \quad X_k = m_i + \epsilon_k,$$

les  $\epsilon_k$  sont centrées indépendantes de variance  $\sigma_i^2$  et  $i$  est la cellule à laquelle appartient  $k$ . Les estimateurs de Gauss-Markov des  $m_i$  sont les moyennes observées dans les différentes cellules  $y_i$ . L'estimateur par prédiction vaut alors:

$$(3.1.2) \quad X^{\text{Pred}} = \sum_i (N_i - n_i) y_i + \sum_i n_i y_i = \sum_i N_i y_i.$$

Il a la forme de l'estimateur poststratifié. On obtient de plus, immédiatement que:

$$(3.1.3) \quad \text{Var}(X^{\text{Pred}} - Y)^2 = \sum_i \sigma_i^2 N_i (N_i - n_i) / n_i.$$

Cette quantité ne dépend pas de l'échantillon  $s$  puisque celui-ci comporte toujours (avec probabilité 1)  $n_i$  individus de la cellule  $i$ . L'estimation de  $E(X^{\text{Pred}} - Y)^2$  se fait en remplaçant  $\sigma_i^2$  par son estimateur habituel  $s_i^2 = (n_i - 1)^{-1} \sum_{k \in s_i} (X_k - y_i)^2$  avec  $s_i$  partie de  $s$  dans la cellule  $i$ . Ces résultats sont dus à Gouriéroux (1981) et constituent, dans une certaine mesure, une justification de la méthode des quotas simples.

formé par les quantités:  $N_{i++\dots+h}^{\dots}$ ,  $N_{+f+\dots+h}^{\dots}$ ,  $N_{++\dots+h}^{\dots}$  pour  $i = 1 \text{ à } I$ ,  $j = 1 \text{ à } J - 1$ , et  $h = 1 \text{ à } H - 1$  (de façon à ne conserver que des variables linéairement indépendantes). Les régresseurs sont alors les variables indicatrices des catégories  $i$  ( $i = 1 \text{ à } I$ ),  $j$  ( $j = 1 \text{ à } J - 1$ ) et  $h = 1 \text{ à } H - 1$ ). Comme la constante est combinaison linéaire des régresseurs (c'est la somme de  $I$  premiers d'entre eux) l'estimateur par régression prendra la forme:

$$Y_{\text{Reg}} = \sum_{i=1}^I N_{i++\dots+h}^{\dots} A_i + \sum_{j=1}^J N_{+f+\dots+h}^{\dots} B_j + \dots + \sum_{h=1}^H N_{++\dots+h}^{\dots} C_h, \quad (2.2.2)$$

où  $A_i$  (par exemple) est le coefficient de l'indicatrice d'appartenance à la catégorie  $i$ .

Si on ne travaille qu'avec une seule catégorisation les régresseurs sont 2 à 2 orthogonaux et on a:

$$Y_{\text{Reg}} = \sum_{i=1}^I N_i Y_i$$

où  $Y_i$  est l'estimateur de la moyenne de  $Y$  dans la catégorie  $i$ .  $Y_{\text{Reg}}$  n'est alors pas autre chose que l'estimateur poststratifié.

### 2.3 Théorie des sondages basée sur des modèles

Dans cette optique, on considère les  $X_k$  comme des variables aléatoires régies par un modèle de superpopulation. Celui-ci comporte des paramètres qu'on estime à partir de l'échantillon. On peut alors calculer l'espérance, sous le modèle estimé, des valeurs non observées de  $Y$ , soit  $Y_k$ . L'estimateur par prédiction, somme des valeurs observées et des valeurs prévues est donné par:

$$Y^{\text{Pred}} = \sum_{k=1}^s Y_k + \sum_{k=s+1}^{U-s} Y_k.$$

Si, par exemple, dans un sondage à probabilités égales, le modèle est une régression  $Y_k = X_k \cdot \beta + \epsilon_k$ ,  $\epsilon_k$  indépendantes, centrées, d'égales variances et que la constante figure dans la régression (ou qu'une combinaison linéaire de  $X_k$  soit constante) alors on a  $\sum_s X_k \cdot \beta = \sum_s Y_k$  et l'estimateur par la prédiction se confond avec l'estimateur par la régression. On dira que  $Y$  est sans biais sous le modèle si, pour tout  $s$ ,  $E(Y - Y^{\text{Pred}}) = 0$  (on note  $E$  et  $V$  l'espérance et la variance sous le modèle, conditionnellement à l'échantillon). Pour l'estimateur par prédiction il suffit qu'on ait pour tout  $k$  la condition naturelle  $E Y_k = E X_k$  pour que cela soit vrai. Sous le modèle on peut évaluer également l'écart quadratique moyen:  $E(Y^{\text{Pred}} - Y)^2$  sachant que les deux termes  $Y^{\text{Pred}}$  et  $Y$  sont aléatoires et que  $Y^{\text{Pred}}$  dépend de l'échantillon  $s$ . L'espérance ci-dessus est donc une espérance conditionnelle à l'échantillon  $s$ . Celui-ci a une certaine loi de probabilité déjà introduite au paragraphe précédent. On peut mesurer la précision de cet estimateur en calculant:

$$V(Y^{\text{Pred}}) = E E(Y^{\text{Pred}} - Y)^2.$$

Si la loi de  $s$  et celle des  $Y_k$  sont indépendantes (échantillonnage dit non informatif) alors cette quantité est égale à:

$$E(E(Y^{\text{Pred}} - Y)^2),$$

où l'espérance intérieure est prise conditionnelle aux  $Y_k$ . Si  $Y^{\text{Pred}}$  s'identifie à  $Y_{\text{Reg}}$ , et que la condition d'indépendance est réalisée, on aura donc:

$$V(Y^{\text{Pred}}) = E(\text{Var}(Y_{\text{Reg}})).$$

Cet estimateur a une variance connue qui est une forme quadratique  $V(Y_U)$  sur le vecteur des  $Y_k$  dans la population:

$$\text{Var}(Y) = V(Y_U) = V(d_k - 1) + \sum_{kl} Y_k Y_l d_k d_l (\pi_{kl} - \pi_k \pi_l), \quad (2.2.1)$$

où  $\pi_{kl}$  est la probabilité d'avoir simultanément  $k$  et  $l$  dans  $s$ . De même on peut estimer sur les données la variance de  $Y$  par une forme quadratique formée sur le vecteur  $Y_s$  des  $Y_k$  de l'échantillon:

$$V(Y_s) = \sum_{k,l \in s} \Delta_{kl} Y_k Y_l,$$

avec

$$\Delta_{kl} = (1 - \pi_k) / \pi_k^2 \quad \text{si } k = l$$

$$= (\pi_{kl} - \pi_k \pi_l) / (\pi_{kl} \pi_k \pi_l) \quad \text{si } k \neq l.$$

Selon les plans de sondage ces expressions prennent des formes particulières qu'on trouve dans les manuels (Desabie 1965, Cochran 1977, Wolter 1985).  
Toute information externe peut améliorer la qualité de l'estimation. Celle-ci, souvent, se présente sous la forme d'un vecteur  $X$  dont chacune des  $p$  composantes est le total d'une variable mesurable dans chacun des échantillons possibles. On peut alors améliorer l'estimation de  $Y$  en utilisant l'estimation par régression:

$$Y_{\text{reg}} = Y + (X - \bar{X})' \bar{b},$$

où  $\bar{b}$  est le vecteur des coefficients de la régression des  $Y_k$  sur les  $X_k$  estimé par:

$$\bar{b} = \sum_{k=1}^s (d_k X_k X_k')^{-1} \sum_{k=1}^s d_k X_k Y_k.$$

Dans le cas où la constante fait partie des régresseurs où si elle est combinaison linéaire des régresseurs et que l'échantillon est à probabilités égales la formule se simplifie en:

$$Y_{\text{reg}} = X' \bar{b}.$$

La variance de  $Y_{\text{reg}}$  s'exprime simplement en introduisant les résidus de la régression  $E_k = Y_k - X_k' \bar{b}$  dans la population. Il est connu qu'on a:

$$\text{Var}(Y_{\text{reg}}) = V(E_U)$$

où on porte donc dans la formule (2.2.1) le vecteur  $E_U$  des résidus  $E_k$ . De même on estime approximativement cette variance par  $V(e_s)$  où  $e_s$  est le vecteur des  $e_k = Y_k - X_k' \bar{b}$ , résidus estimés de la régression.

Sous divers plans de sondage ces expressions prennent des allures particulières. En règle générale,  $V$  et  $V'$  étant des formes quadratiques positives et les  $E_k$  ou  $e_k$  des quantités plus petites que les  $Y_k$ , l'estimateur par régression conduit à des améliorations substantielles par rapport aux valeurs diluées.

Un cas particulier important que nous utiliserons dans la suite est celui où  $X$  est un vecteur de totaux de variables de comptage (effectifs à partir desquels on construit des quotas). Typiquement l'information auxiliaire est le vecteur de dimension  $I + (J - 1) + \dots + (H - 1)$



2. UNE REVUE RAPIDE DE LA MÉTHODE DES QUOTAS  
ET DE LA THÉORIE DES SONDAGES

2.1 Quotas sur des cellules; quotas sur les marges d'une table de contingence – quelques aspects pratiques de la méthode

Au niveau le plus simple, la méthode des quotas ressemble à l'échantillonnage stratifié. On connaît la répartition dans la population d'un caractère discret  $h$  que  $N_h$  individus possèdent ( $h = 1 \text{ à } H$ ). L'échantillon comporte  $n_h$  individus de catégorie  $h$  dont le choix est laissé aux enquêteurs. Le taux de sondage  $f_h = n_h/N_h$  peut, éventuellement, varier de catégorie en catégorie. En pratique, on préfère, généralement, contrôler plusieurs critères qu'on notera  $i, j, \dots, h$  ( $i = 1 \text{ à } I, j = 1 \text{ à } J, \dots, h = 1 \text{ à } H$ ). Idéalement, la connaissance des effectifs  $N_{ij\dots h}$  du tableau de contingence à entrées multiples permettrait de se ramener à la méthode précédente pour la définition d'effectifs  $n_{ij\dots h}$  composant l'échantillon selon des taux  $f_{ij\dots h}$ . Sauf dans des cas très particuliers (peu de critères ayant chacun peu de modalités) cette méthode est irréalisable car elle conduit à la recherche d'individus extrêmement difficiles à détecter.

On préfère fixer des **quotas marginaux** en calibrant l'échantillon de façon à ce que sa répartition selon le premier critère conduise à des effectifs  $n_{i+,\dots,+}$  donnés, de même en ce qui concerne les autres critères. La seule contrainte sur ces effectifs marginaux est de s'additionner à  $n$ , effectif global de l'échantillon. Pratiquement, on adopte presque toujours un taux de sondage  $f$  unique pour chaque batterie de quotas:  $n_{i+,\dots,+} = fN_{i+,\dots,+}$ ,  $n_{+j,\dots,+} = fN_{+j,\dots,+}$  et  $n_{+,+,\dots,h} = fN_{+,+,\dots,h}$  avec des notations évidentes (+ à la place d'un indice indiquant la sommation sur toutes les modalités de la catégorie notée par l'indice).

Outre son avantage évident de collecte, cette technique est la plus souvent imposée par les données externes sur lesquelles on assoie les quotas. Celles-ci proviennent, par exemple, de sources différentes interdisant tout croisement. Une autre situation se présente quand les quotas sont établis à partir d'une grosse enquête (par exemple sur l'emploi): chaque répartition selon un critère (âge, catégorie socio-professionnelle *etc.*) peut être considérée comme fiable. En revanche, les croisements sont entachés d'une erreur aléatoire importante et ne peuvent pas être utilisés pour fixer des quotas.

En pratique, la méthode des quotas est la plus souvent utilisée en complément de méthodes plus traditionnelles comme ultime technique d'échantillonnage dans une enquête stratifiée à plusieurs degrés sur des bases géographiques (région, taille des agglomérations). Chaque unité primaire est confiée à un enquêteur à qui sont fixés des quotas. Celui-ci reçoit, de plus, des consignes destinées à disperser son échantillon de façon à rapprocher le recueil des données de ce qu'aurait donné le hasard.

2.2 La théorie traditionnelle des sondages

On désire mesurer le total  $Y$  d'une variable dont la valeur  $Y_k$  pour l'individu  $k$  n'a rien d'aléatoire. Seul l'échantillon  $s$  est aléatoire et sa loi de probabilité est connue car contrôlée par le statisticien. Par suite, la probabilité  $\pi_k$  qu'un individu d'appartenance dans  $s$  est aussi connue. Sans autre information, l'estimateur naturel (sans biais) à retenir est l'estimateur par les valeurs latentes:

$$Y = \sum_{k \in s} Y_k / \pi_k = \sum_{k \in s} d_k Y_k \text{ avec } d_k = 1 / \pi_k.$$

Dans le cas où les  $\pi_k$  sont tous égaux à  $n/N$ , le taux de sondage, on a:

$$Y = N/n \sum_{k \in s} Y_k = N\bar{y},$$

où  $\bar{y}$  désigne la moyenne de  $Y$  sur l'échantillon.

# Une théorie des enquêtes par quotas

JEAN-CLAUDE DEVILLE<sup>1</sup>

## RÉSUMÉ

Les enquêtes par quotas simples ou marginaux sont analysées par deux méthodes: (1) modélisation des comportements (modèle de superpopulation) et estimation par prédiction et (2) modélisation de l'échantillonnage (sondage aléatoire simple sous contraintes) et estimation dérivée de la distribution échantillonnale. Dans les deux cas on précise les limites de la théorie, à l'interieur de laquelle on établit des formules de variance et d'estimation de variance quand on mesure des totaux. Une extension de la méthode des quotas (quotas non-proportionnels) est, au passage, décrite et analysée. Elle autorise, dans certains cas, une très nette amélioration de la précision des enquêtes. Les mérites de la méthode des quotas sont comparés à ceux de l'échantillonnage aléatoire. Ce dernier reste indispensable dans le cas d'enquêtes de grande taille dans le cadre de la Statistique officielle.

**MOTS CLÉS:** Enquêtes par quotas; modèles de superpopulation; échantillonnage contraint; estimation par régression.

## 1. INTRODUCTION

L'échantillonnage par quotas est la méthode la plus fréquemment utilisée en France par les Instituts de sondage privés. Facile de mise en oeuvre, peu coûteuse, elle possède de nombreux avantages pratiques. Ses défauts, toutefois, sont aussi assez bien connus: possibilités de biais, impossibilité de traiter les non-réponses, nécessité d'une information externe pour fixer les quotas. Dans la littérature anglo-saxonne (Cochran 1977 ou Madow et coll. 1983 par exemple) les quotas ont fort mauvaise réputation à cause de l'absence d'une théorie fiable sur laquelle une inférence statistique puisse être fondée. Les "défenseurs" de la méthode (Smith 1983 en particulier) se basent sur des principes d'inférence conditionnelle à l'échantillon où le plan de sondage peut, généralement, être oublié.

Cet article propose une théorie des enquêtes par quotas basée sur deux types de modélisation: modélisation du comportement de la population (qui est l'optique de Smith ou des idées exprimées dans Gouriéroux 1981), modélisation du mode de recueil de l'échantillon, ce qui correspond, peut-être, à une idée plus réaliste.

Dans chaque cas, on obtient les variances des estimateurs en se ramenant des variantes d'estimateurs par régression.

L'article commence par une description de la méthode des quotas et des résultats de théorie des sondages utiles pour la suite. Les parties 2 et 3 développent des modèles de comportement des individus de la population ou des enquêteurs, qui justifient la méthode. La dernière partie évoque des problèmes ouverts et montre en quoi la méthode des quotas complète les méthodes probabilistes classiques plus qu'elle ne les concurrence.

<sup>1</sup> Jean-Claude Deville, Institut National de la Statistique et des Etudes Economiques, 18, Boulevard Adolphe Pinard, 75675, Paris Cedex 14, France.

SUTRADHAR, B.C., et MACNEILL, I.B. (1989). Two-way analysis of variance for stationary periodic time series. *Revue Internationale de Statistique*, 57, 169-182.

SUTRADHAR, B.C., MACNEILL, I.B., et DAGUM, E.B. (1991). A Simple Test for Stable Seasonalities. Statistique Canada, Document de travail de la Direction de la méthodologie, N° TSRA-91-007.

SUTRADHAR, B.C., MACNEILL, I.B., et SAHRMANN, H.F. (1987). Time series valued experimental designs: One-Way Analysis of Variance with Autocorrelated Errors. Dans *Time Series and Econometric Modelling*, (Eds. I.B. MacNeill et G.J. Umphrey) Dordrecht: Reidel, 113-129.



mobiles saisonnier (MMS) dans lequel  $\theta$  ou  $\Theta$  les deux sont significatifs. Le test  $F$  exact donne des résultats très différents de ceux du test  $F$  de la X-11-ARMMI (ou de la variante X-11) lorsque l'autocorrélation des résidus est saisonnière, c.-à-d. lorsque s'écarte de zéro de façon significative.

En ce qui a trait aux séries mensuelles analysées, le test  $F$  ordinaire et le test  $F$  modifié aboutissent à des résultats contradictoires dans deux cas seulement sur vingt-six. En revanche, si nous appliquons la règle empirique du  $F \geq 7$  pour justifier la désaisonnalisation, le test  $F$  modifié produirait des résultats contradictoires dans huit cas sur douze.

En examinant les valeurs désaisonnalisées des huit séries en question, nous avons constaté que six d'entre elles pouvaient être désaisonnalisées convenablement à l'aide de la méthode X-11-ARMMI.

Pour ce qui est des séries trimestrielles, le test  $F$  modifié indique que deux des quatre séries analysées ne renferment pas de mouvement saisonnier stable. Par ailleurs, dans un cas, le test  $F$  de la X-11-ARMMI produit une valeur supérieure à 7 alors que le test  $F$  modifié entraîne l'acceptation de l'hypothèse nulle.

Nous avons supposé dans cet article que le seul type de mouvement saisonnier évolutif qui pouvait être présent dans les séries était des variations annuelles. Le test que nous venons de décrire n'est pas conçu pour déceler d'autres types de mouvement saisonnier évolutif. Il y a donc lieu de poursuivre la recherche sur ce plan.

## REMERCIEMENTS

Les auteurs tiennent à remercier l'arbitre anonyme qui a exprimé de précieux commentaires sur une version antérieure du document.

## BIBLIOGRAPHIE

- BOX, G.E.P., et JENKINS, G.M. (1970). *Time Series Analysis, Forecasting and Control*. San Francisco: Holden-Day.
- DAGUM, E.B. (1980). *La méthode de désaisonnalisation X-11-ARMMI*. N° 12-564F au catalogue, Statistique Canada.
- FRANZINI, L., et HARVEY, A.C. (1983). Testing for deterministic trend and seasonal components in time series models. *Biometrika*, 70, 673-682.
- PIERCE, D.A. (1978). Seasonal adjustment when both deterministic and stochastic seasonality are present. Dans *Seasonal Analysis of Economic Time Series*, (Ed. A. Zellner). Washington, D.C.: U.S. Bureau of the Census, 242-272.
- SATTERTHWAITE, F.E. (1946). An approximate distribution of estimates of variance components. *Biometrics*, 2, 110-114.
- SHISKIN, J., YOUNG, A.H., et MUSGRAVE, J.C. (1967). The X-11 Variant of Census Method II: Seasonal Adjustment Program. Technical Paper 15, Bureau of the Census, U.S. Dept. of Commerce.
- SHISKIN, J., et FLEWES, T. (1978). Seasonal adjustment of the U.S. unemployment rate. *The Statistician*, 27, 181-202.
- SUTRADHAR, B.C., et BARTLETT, R.F. (1989). An approximation to the distribution of the ratio of two general quadratic forms with application to time series valued designs. *Communications in Statistics - Theory and Methods*, 18, 1563-1588.
- SUTRADHAR, B.C., et BARTLETT, R.F. (1990). An Exact Large and Small Sample Comparison of Wald's, Likelihood Ratio and Rao's Tests for Testing Linear Regression with Autocorrelated Errors. Rapport technique, Department of Mathematics and Statistics, Memorial University of Newfoundland.

Tableau 2

Détermination de la présence d'un mouvement saisonnier stable dans des séries trimestrielles					
Séries	Valeur estimées des paramètres		Test $F^a$ X-11-ARMMI	$F_{M2}$ (valeur $p$ en %)	Décision finale <sup>c</sup>
	$\theta_1$	$\theta_1$			
1. Dépôts dans les autres institutions	0.53*	0.11*	9.03	9.67(0.04)	O
2. Investissement financier net	0.77*	-0.37*	4.86 <sup>b</sup>	2.56(8.16)	N
3. Petites hypothèques	0.17*	-0.01	6.65	4.88(1.02)	O
4. Créances d'entreprises privées	0.77*	-0.31*	7.88 <sup>b</sup>	3.58(3.20)	N

a La valeur critique est  $F(3,27; 0.01) = 4.51$ .  
b Résultat contradictoire à celui du test  $F$  modifié.  
c O (Oui) - mouvement saisonnier stable significatif.  
N (Non) - absence de mouvement saisonnier stable.  
\* Valeur significative à un seuil de 5%.

les quatre séries. Le test  $F$  modifié s'exécute à peu près de la même manière que pour les séries mensuelles, mais comme la matrice des covariances  $\Sigma^*$  est différente, il a fallu modifier en conséquence les formules de  $C_1(\cdot)$ ,  $C_2(\cdot)$ , et  $C_3(\cdot)$  dans les équations (2.6) et (2.7). Comme dans le cas des séries mensuelles, nous avons constaté que les valeurs  $p$  qui servent à tester la présence de variations annuelles (test  $F_{M1}$  modifié) étaient très élevées, ce qui nous a amenés à exclure la possibilité d'un mouvement saisonnier évolutif sous forme de variations annuelles.

Quant au test de la présence d'un mouvement saisonnier stable dans chacune des quatre séries, nous donnons les résultats du test  $F_{M2}$  modifié et du test  $F$  de la X-11-ARMMI dans le tableau 2. La valeur  $p$  pour deux de ces séries, notamment "Dépôts dans les autres institutions" et "Petites hypothèques", n'est pas significative et se compare à celles obtenues par la X-11-ARMMI. Nous devons en conclure que ces deux séries renferment un mouvement saisonnier stable significatif. Pour ce qui est des deux autres séries trimestrielles, le test  $F$  modifié et le test  $F$  de la X-11-ARMMI donnent des résultats contradictoires. Le premier, contrairement au second, produit des valeurs  $p$  significatives pour les deux séries. En conclusion, les séries "Investissement financier net" et "Créances d'entreprises privées" ne devraient pas être désaisonnalisées.

4. CONCLUSIONS

Dans cet article, nous avons proposé un test exact pour vérifier la présence d'un mouvement saisonnier stable et de variations saisonnières annuelles dans une série. Inspiré du test  $F$  modifié de Sutrathdar, MacNeill et Sahrmann (1987), le nouveau test tient compte de l'auto-corrélation des résidus dans les rapports "composante saisonnière-composante irrégulière" de la méthode X-11-ARMMI. On suppose que les résidus suivent un processus à moyennes mobiles saisonnier (MMS) simple  $(0,q)(0,Q)$ . Nous avons appliqué ce test à des séries trimestrielles tirées du Système de comptabilité nationale ainsi qu'à un ensemble de séries mensuelles portant sur les importations, les exportations, les prix à la consommation et la main-d'oeuvre. Nous avons observé que les résidus de la méthode X-11-ARMMI suivent un modèle à moyennes



Comme le test  $F$  modifié est inefficace lorsqu'il existe un mouvement saisonnier évolutif (à l'exception des variations saisonnières annuelles), nous nous sommes assurés à l'aide de tests préliminaires tirés de la X-11-ARMMI que les séries étudiées ne renfermaient aucun mouvement saisonnier évolutif. (Nous avons aussi examiné la représentation graphique des rapports "composante saisonnière-composante irrégulière".)

La méthode X-11-ARMMI a servi à la décomposition des séries  $\{Z_t: t = 1, \dots, 120\}$ . Des tests de diagnostic montrent que la composante d'erreur  $U_t$  (voir équation 2.1) suit un processus SARMA  $(0,1)(0,1)^{12}$  pour chacune des séries. Les estimations  $\theta_1$  et  $\Theta_1$  entrent dans le calcul des statistiques  $F_{M1}$  modifiées et  $F_{M2}$ .

En ce qui a trait au test visant à vérifier la présence de variations saisonnières annuelles (test  $F$  modifié), nous avons constaté que l'approximation de Satterthwaite et le test  $F$  (d'analyse de variance ordinaire donnaient, en règle générale, des valeurs  $p$  différentes. Dans les deux cas toutefois, ces valeurs étaient très élevées pour chacune des séries, indiquant par le fait même l'absence d'un mouvement saisonnier évolutif sous forme de variations annuelles.

Pour ce qui a trait au test de la présence d'un mouvement saisonnier stable, nous avons calculé, pour les 26 séries mensuelles, les valeurs  $p$  de la statistique  $F$  modifiée  $F_{M2}$  (équ. 2.7) au moyen de l'approximation de Satterthwaite et avons comparé ces valeurs à celles obtenues à l'aide du test  $F$  de la X-11-ARMMI (qui est l'équivalent du test  $F$  de l'analyse de variance ordinaire). Les résultats pertinents figurent dans le tableau 1.

D'après les valeurs  $p$  de la statistique  $F$  modifiée, trois des neuf séries sur les importations ne renferment pas de mouvement saisonnier stable significatif à un seuil de 1% (valeur critique de  $F(11,99; 0,01) = 2,47$ ). Parmi les sept séries sur les exportations, une seule – Bile – semble ne pas présenter de mouvement saisonnier. Quant aux séries sur l'IPC, les six affichent un mouvement saisonnier stable; il en va de même pour les quatre séries sur la main-d'oeuvre.

Le test  $F$  de la X-11-ARMMI donne des résultats semblables à ceux du test  $F$  modifié (rejet ou acceptation de l'hypothèse nulle) pour un grand nombre de séries. Il semble que pour la plupart des séries mensuelles, suivant une structure d'erreur SARMA  $(0,1)(0,1)^s$ , le test  $F$  de la X-11-ARMMI (ou, si l'on veut, le test  $F$  de l'analyse de variance ordinaire) est plus sensible aux valeurs négatives élevées de  $\Theta_1$ , c.-à-d. qu'il donne des résultats sensiblement différents de ceux du test  $F$  modifié lorsqu'il existe une autocorrélation saisonnière entre les résidus. Nous pouvons généraliser cette observation en examinant les valeurs de  $C_3(\theta, \Theta)/C_2(\theta, \Theta)$ . En vérifiant dans quelles circonstances ce rapport prend une valeur supérieure à  $\theta_1$  ou une valeur inférieure à 1, on remarque que le sens de l'inégalité dépend du signe de  $\theta_1$  et que la grandeur du rapport dépend de la valeur de  $\Theta_1$ . Dans deux cas seulement, soit la série sur les importations de moteurs d'avions et celle sur les importations de matériel de transport, le test  $F$  ordinaire et le test  $F$  modifié aboutissent à des résultats contradictoires. En revanche, si nous appliquons la règle empirique du  $F \geq 7$  pour justifier la désaisonnalisation, le test  $F$  modifié donnerait des résultats contradictoires dans huit cas sur douze. Nous avons donc désaisonné les huit séries en question à l'aide de la méthode X-11-ARMMI et avons constaté que la qualité de la correction était bonne dans six cas sur huit. Toutes les séries ont été soumises au modèle d'extrapolation ARMMI qui avait été choisi automatiquement pour le programme, six des huit séries traitées ont satisfait aux critères d'acceptation de la X-11-ARMMI, et les quatre séries pour lesquelles la valeur de  $F_{M2}$  était relativement petite, c'est-à-dire de 3,24 à 3,74, étaient nettement influencées par les variations des jours ouvrables. Dans deux cas seulement, soit la série sur les importations de fourrages et de provendes et celle sur les importations de produits végétaux bruts, nous avons constaté que les valeurs désaisonnalisées ne pouvaient être jugées fiables.

### 3.2 Séries trimestrielles

Nous avons appliqué la méthode X-11-ARMMI à quatre séries trimestrielles du Système de comptabilité nationale afin d'obtenir les valeurs de la série décomposée ( $Z_t, t = 1, \dots, 40$ ). Nous avons pu ainsi constater que la composante d'erreur  $U_t$  suit un modèle  $(0,1)(0,1)^4$  pour



Tableau 1  
Détermination de la présence d'un mouvement saisonnier stable dans des séries mensuelles

Séries	Valeurs estimées des paramètres		Test $F^a$	Test $F$ modifié $F_{m2}$ (valeur $p$ en %)	Décision finale <sup>c</sup>
	$\theta_1$	$\theta_1$			

IMPORTATIONS					
1. Fourrages et provenances	-0.09*	-0.01	3.68	3.43(0.06)	O
2. Houille et produits connexes	0.02	-0.01	64.40	58.76(0.00)	O
3. Produits végétaux bruts	0.02	-0.07*	3.48	2.94(0.27)	O
4. Laine et fibres artificielles	0.02	0.29*	10.98	20.63(0.00)	O
5. Métaux précieux	0.27*	0.01	1.25	1.20(31.10)	N
6. Huiles et matières grasses	0.41*	0.01	8.59	8.22(0.00)	O
7. Minéraux non métalliques	0.04	0.02	16.50	16.68(0.00)	O
8. Moteurs d'avions	0.32*	0.00	2.53 <sup>b</sup>	2.36(1.79)	N
9. Autres matériel de transport	0.19*	-0.18*	3.48 <sup>b</sup>	2.43(1.31)	N

EXPORTATIONS					
10. Blé	0.04	-0.03	1.89	1.71(8.71)	N
11. Armiante	0.13*	-0.03	6.83	6.15(0.00)	O
12. Pâte de bois	-0.27	0.20*	6.45	9.61(0.00)	O
13. Demi-produits en matières textiles	0.52*	0.13*	12.05	15.06(0.00)	O
14. Autres matières travaillées, non comestibles	0.04	0.11*	5.03	6.19(0.00)	O
15. Téléviseurs, équipement de télécommunication	0.12*	0.01	9.26	8.99(0.00)	O
16. Voitures particulières	-0.30*	-0.14*	24.50	18.52(0.00)	O

IPC					
17. Oeufs	-0.04	-0.01	6.90	6.50(0.00)	O
18. Pâtes	-0.05*	-0.04	3.69	3.24(0.10)	O
19. Oignons	-0.42*	-0.03	26.90	23.49(0.00)	O
20. Logement	0.11*	-0.34*	19.02	9.28(0.00)	O
21. Habillement	0.03	-0.42*	47.42	24.30(0.00)	O
22. Transport	-0.09*	-0.02	4.21	3.74(0.02)	O

MAIN-D'OEUVRE					
23. Personnes occupées (25-34), Saskatchewan	-0.19*	-0.11*	67.40	52.35(0.00)	O
24. Personnes inactives, Saskatchewan	0.12*	-0.36*	22.98	12.69(0.00)	O
25. Personnes en chômage (25-44), Ontario	-0.21*	0.07*	31.4	34.23(0.00)	O
26. Personnes en chômage (20-24), Ontario	-0.02	0.19*	24.27	34.78(0.00)	O

a La valeur critique est  $F(11,99; 0.01) = 2.47$ .  
b Résultat contradictoire à celui du test  $F$  modifié.  
c O (Oui) - mouvement saisonnier stable significatif.  
N (Non) - absence de mouvement saisonnier stable.  
\* Valeur significative à un seuil de 5%.

2.3 Calcul de la valeur  $p$

Une étude de simulation (voir Sutradhar et Bartlett 1989, tableau IV, p. 1587) montre que lorsqu'il y a  $k$  groupes indépendants, la distribution  $F$  habituelle peut, dans le cas d'un processus SARMA SMA/(0,q) (0,Q)<sub>s</sub>, servir d'approximation pour la distribution de la statistique  $F$  modifiée. En règle générale, toutefois, une telle approximation est incorrecte, surtout lorsque les  $k$  groupes sont corrélés et que  $n$  est petit.

Dans cet article, nous servons de l'approximation bien connue de Satterthwaite (1946) (voir Sutradhar, MacNeill et Dagum 1991) pour calculer les valeurs  $p$ , notamment  $P_r(F_{M1} \geq f_{M1})$ , où  $f_{M1}$  est la valeur empirique de  $F_{M1}$ . À cette fin, nous calculons tout d'abord les valeurs propres  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_r > 0 = \lambda_{r+1} = \dots = \lambda_s > \lambda_{s+1} \geq \dots \geq \lambda_n$  de

(2.8) 
$$\Sigma^{*1/2} [d_1(\theta, \theta) D_1 - f_{M1} (I^{kn} - D_2)] \Sigma^{*1/2},$$

où  $d_1(\cdot)$  est défini dans l'équation (2.6),  $D_1 = R(RR')^{-1}R'$ , où  $R = C(X'X)^{-1}$ ,  $D_2 = X(X'X)^{-1}X'$ ,  $C$ . Dans l'expression ci-dessus,  $C$  est une matrice appropriée que l'on obtient en exprimant l'hypothèse nulle  $H_0: \beta_f = 0$  sous la forme  $C\gamma = 0$ , où  $\gamma$  est défini en (2.5). Dans l'équation (2.8),  $I^{kn}$  est la matrice unité  $kn \times kn$ . L'approximation de Satterthwaite donne donc le résultat suivant:

(2.9) 
$$P_r(F_{M1} \geq f_{M1}) = P_r[F_{a,b} \geq bd/ac],$$

où  $F_{a,b}$  désigne le ratio  $F$  habituel avec les degrés de liberté  $a$  et  $b$ , où

$$a = \left( \sum_{f=1}^r \lambda_f \right)^2 / \sum_{f=1}^r \lambda_f^2, \quad b = \left( \sum_{n}^{f=s+1} \lambda_f \right)^2 / \sum_{n}^{f=s+1} \lambda_f^2.$$

Dans l'équation (2.9),

$$c = \sum_{f=1}^r \lambda_f^2 / \sum_{f=1}^r \lambda_f, \quad d = \sum_{n}^{f=s+1} \lambda_f^2 / \sum_{n}^{f=s+1} |\lambda_f|.$$

De même, on peut calculer  $P_r(F_{M2} \geq f_{M2})$  en substituant  $d_2(\cdot)$  et  $f_{M2}$  à  $d_1(\cdot)$  et à  $f_{M1}$  respectivement dans l'équation (2.9). La construction de  $D_1$  dépendra alors d'une matrice  $C$  différente, que l'on obtiendra en exprimant l'hypothèse nulle  $H_0: \alpha_f = 0$  sous la forme  $C\gamma = 0$ .

3. APPLICATIONS

3.1 Séries mensuelles

Nous avons calculé la valeur des statistiques  $F$  modifiées  $F_{M1}$  et  $F_{M2}$  (équations 2.6 et 2.7) pour 26 séries mensuelles touchant divers secteurs économiques, notamment les importations, les exportations, les prix à la consommation et la main-d'oeuvre. Toutes les séries portent sur la période de janvier 1979 à décembre 1988 inclusivement.

avec

$$\bar{Z}_i = \sum_n^j Z_{ij}/n, \quad \bar{Z}_j = \sum_k^i Z_{ij}/k, \quad \text{et} \quad \bar{Z}_{..} = \sum_n^i \sum_k^j Z_{ij}/kn,$$

$Z_{ij}$  étant la  $j$ -ième observation dans la  $i$ -ième période. Or, dans le cas qui nous occupe, ces statistiques ne conviennent pas pour tester les hypothèses mentionnées ci-dessus parce que l'espérance des sommes des carrés est influencée par l'interdépendance des observations. De plus, les sommes des carrés ne sont pas indépendantes les unes des autres. Lorsque  $U^*$  dans l'équation (2.5) suit un processus SARMA  $(0, q) (0, \bar{Q})^s$ , il est possible de montrer que

$$E(\bar{Q}_1) = k \sum_n^j \beta_j^2 + \sigma^2(n-1)C_1(\theta, \theta),$$

$$E(\bar{Q}_2) = n \sum_k^i \alpha_i^2 + \sigma^2(k-1)C_2(\theta, \theta),$$

et

$$E(\bar{Q}_3) = \sigma^2(k-1)(n-1)C_3(\theta, \theta),$$

où, par exemple dans le cas d'un processus SARMA  $(0, 1) (0, 1)^2$ ,

$$C_1(\theta, \theta) = (1 + \theta_1^2)(1 + \theta_2^2) - (\theta_1/6)(1 + \theta_1^2)(1 + \theta_2^2) + (2\theta_1/n)(1 + \theta_1^2) + (\theta_1\theta_2/6)\{1 - 22/n - (n-2)/n(n-1)\},$$

$$C_2(\theta, \theta) = (1 + \theta_1^2)(1 + \theta_2^2) - 2(1 - 1/n)\theta_1(1 + \theta_1^2)$$

$$+ 1/6\{1 + (1 - 1/n)/11\}\theta_1(1 + \theta_1^2) - (4/11)(1 - 1/n)\theta_1\theta_2,$$

$$C_3(\theta, \theta) = (1 + \theta_1^2)(1 + \theta_2^2) + (2\theta_1/n)(1 + \theta_1^2) + (\theta_1/6)(1 + \theta_1^2)(1 + \theta_2^2)(1 - 1/11n)$$

$$- (\theta_1\theta_2/6n)[n/11 - 2(n-2)/11(n-1) - 2].$$

Par conséquent, on peut tester les hypothèses nulles  $\beta_j = 0$  et  $\alpha_i = 0$  en se servant des statistiques  $F$  modifiées  $F_{M1}$  et  $F_{M2}$  respectivement, celles-ci étant définies

$$(2.6) \quad F_{M1} = d_1(\theta, \theta)F_{A1},$$

$$(2.7) \quad F_{M2} = d_2(\theta, \theta)F_{A2},$$

(voir aussi Sutrathar, MacNeill et Sahrman 1987; Sutrathar, MacNeill et Dagum 1991), où  $d_1(\theta, \theta) = C_3(\theta, \theta)/C_1(\theta, \theta)$ ,  $d_2(\theta, \theta) = C_3(\theta, \theta)/C_2(\theta, \theta)$ . Ces statistiques tiennent compte de l'autocorrélation des résidus. Notons que dans l'hypothèse où il y a indépendance, c'est-à-dire lorsque  $\theta = 0$ ,  $\theta = 0$ ,  $C_1(\cdot) = C_2(\cdot) = C_3(\cdot) = 1$ , ce qui va de soi. Dans ces conditions, on peut tester les hypothèses au moyen du critère  $F$  de l'analyse de variance ordinaire.



Les  $\alpha$  et les  $\beta$  dans l'équation (2.2) représentent respectivement le mouvement saisonnier stable et les variations saisonnières annuelles dans une série chronologique saisonnière. Ainsi, lorsqu'il s'agit de vérifier la présence d'un mouvement saisonnier stable, nous testons l'hypothèse

$$H_0: \alpha_i = 0 \quad \text{vs} \quad H_1: \alpha_i \neq 0 \text{ pour au moins un } i; \tag{2.3}$$

et lorsqu'il s'agit de vérifier l'existence de variations saisonnières annuelles, nous testons l'hypothèse

$$H_0: \beta_j = 0 \quad \text{vs} \quad H_1: \beta_j \neq 0 \text{ pour au moins un } j. \tag{2.4}$$

Par conséquent, le rejet de l'hypothèse nulle dans (2.3) et dans (2.4) signifierait que la série en question renferme un mouvement saisonnier stable significatif de même que des variations saisonnières annuelles.

Compte tenu du modèle (2.2), le modèle (2.1) peut s'écrire:

$$Z^* = X\gamma + U^*, \tag{2.5}$$

où

$$\begin{aligned} Z^* &= [Z(1), \dots, Z(n), Z(n+1), \dots, Z(kn)]', \\ U^* &= [U(1), \dots, U(n), U(n+1), \dots, U(kn)]', \\ \gamma &= [\mu, \alpha_1, \dots, \alpha_{k-1}, \alpha_k, \beta_1, \dots, \beta_{n-1}, \beta_n]'. \end{aligned}$$

et  $X$  est la matrice de plan appropriée  $kn \times (k+n+1)$ .

## 2.2 Variable à tester

On peut décrire  $U^*$  dans l'équation (2.5) par un processus stationnaire autorégressif à moyennes mobiles saisonnier (SARMA – seasonal autoregressive moving average)  $(p, q)(P, Q)_s$ . Nous avons toutefois constaté que dans la plupart des cas empiriques, un modèle  $(0, q)(0, \bar{Q})_s$  suffit. Désignons par  $\Sigma^*$  la matrice  $kn \times kn$  des covariances de  $U^*$ . Naturellement,  $\Sigma^*$  contiendra  $\theta \equiv (\theta_1, \dots, \theta_q^b)$  et  $\Theta \equiv (\Theta_1, \dots, \Theta_{\bar{Q}})$ , où  $\theta$  et  $\Theta$  sont les paramètres rattachés au processus SARMA  $(0, q)(0, \bar{Q})_s$ . Dans le cas du modèle d'analyse de variance habituel, c'est-à-dire lorsque les éléments de  $U^*$  sont indépendants et identiquement distribués selon  $N(0, \sigma^2)$ , on teste les hypothèses nulles  $\beta_j = 0$  et  $\alpha_i = 0$  au moyen des statistiques  $F$  classiques  $F_{A1}$  et  $F_{A2}$  respectivement; celles-ci sont définies

$$F_{A1} = (k - 1) \bar{Q}_1 / \bar{Q}_3, \quad \text{et} \quad F_{A2} = (n - 1) \bar{Q}_2 / \bar{Q}_3,$$

où

$$\bar{Q}_1 = k \sum_{n=1}^j (Z_j - \bar{Z}_{..})^2, \quad \bar{Q}_2 = n \sum_k^i (\bar{Z}_i - \bar{Z}_{..})^2,$$

et

$$\bar{Q}_3 = \sum_{k=1}^i \sum_{n=1}^j (Z_{ij} - \bar{Z}_i - \bar{Z}_j + \bar{Z}_{..})^2$$

déterminer la signification de certains coefficients de régression dans un modèle linéaire avec erreurs autocorrélées. Lorsque l'autocorrélation des erreurs est forte, le test classique de Wald, le test du rapport des vraisemblances ou encore les tests qui s'inscrivent dans une analyse par les moindres carrés généralisés posent tous des problèmes de convergence (voir Sutrathar et Bartlett 1990). Pierce (1978) a élaboré un test  $F$  qui repose sur des résidus transformés équivalant à un bruit blanc. La transformation proposée par Pierce revient à utiliser l'inverse de la matrice des covariances des erreurs. Or, cet inverse peut ne pas être calculable lorsque l'autocorrélation des erreurs est forte. Sutrathar, MacNeill et Dagum (1991) ont proposé récemment, pour des modèles linéaires, un test  $F$  modifié visant à vérifier la présence d'un mouvement saisonnier stable. Inspiré des travaux de Sutrathar, MacNeill et Sahrmann (1987), ce test tient compte de l'existence d'une autocorrélation entre les résidus. De plus, il n'exige aucune transformation ni aucune inversion de la matrice des covariances des erreurs.

Franzini et Harvey (1983) ont construit des tests exacts pour vérifier l'hypothèse nulle selon laquelle le mouvement saisonnier est variable, par rapport à l'hypothèse alternative selon laquelle ce mouvement est constant. Contrairement à l'approche de Franzini et Harvey, notre méthode suppose que le mouvement saisonnier est stable, probablement à différents niveaux (à cause des variations annuelles), et vérifie la présence d'un mouvement saisonnier stable (significatif).

Dans la plupart des cas empiriques, un modèle d'erreur à moyennes mobiles saisonnier (SMA – seasonal moving average) du type  $(0,q)(0,Q)_s$  suffit. Dans cet article, nous allons simplifier le test exact qu'ont proposé Sutrathar, MacNeill et Dagum (1991) pour les modèles de ce genre. Ce test vise à vérifier la présence d'un mouvement saisonnier stable et de variations saisonnières annuelles dans un certain nombre de séries socio-économiques.

Voici comment est structuré notre article. La section 2 sert à présenter le test exact. Dans la section 3, nous analysons les résultats de l'application du test  $F$  modifié à un ensemble de séries socio-économiques et nous comparons ces résultats aux valeurs obtenues avec la méthode X-11-ARMMI. Enfin, la section 4 renferme les conclusions.

## 2. TEST F MODIFIÉ

### 2.1 Choix du modèle

Considérons une série chronologique saisonnière stationnaire  $\{Z_t\}$ , définie par l'équation

$$Z_t = S_t + U_t, \tag{2.1}$$

où  $Z_t$  est la série observée au temps  $t$ ,  $S_t$  la composante saisonnière et  $U_t$  les aléas. Si la série renferme une tendance, ce qui est plus que probable, on suppose que par une technique de décomposition appropriée, on obtiendra le modèle (2.1). On peut déduire la série décomposée de la série originale en appliquant à cette dernière les différences appropriées, comme cela se fait avec les modèles ARMMI (Box et Jenkins 1970) ou comme le font depuis longtemps les organismes de statistique en se servant de la méthode X-11-ARMMI ou de la variante X-11 du programme *Census Method II*.

Supposons maintenant qu'il y a  $k$  périodes dans une année et  $kn$  observations dans une série chronologique de  $n$  années. Posons  $Z\{(i-1)n+j\}$  comme la  $j$ -ième ( $j = 1, \dots, n$ ) observation dans la  $i$ -ième période ( $i = 1, \dots, k$ ), qui correspond à  $Z_t$  dans l'équation (2.1). Nous allons définir de la même manière les composantes  $(i,j)$  de  $S_t$  et de  $U_t$ , pour tous  $t = 1, \dots, kn$ . Le modèle supposé pour  $S_t$  est donc (voir Sutrathar et MacNeill 1989):

$$S((i-1)n+j) = \mu + \alpha_i + \beta_j, \tag{2.2}$$

$$\text{où } \sum_{i=1}^k \alpha_i = 0, \sum_{j=1}^n \beta_j = 0.$$

# Test exact pour vérifier la présence d'un mouvement saisonnier stable et applications

BRAJENDRA C. SUTRADHAR, ESTELA BEE DAGUM  
et BINYAM SOLOMON<sup>1</sup>

## RÉSUMÉ

La méthode de désaisonnalisation X-11-ARMMI de même que la variante X-11 du programme *Census Method II* utilisent un test  $F$  d'analyse de variance ordinaire pour déterminer la présence d'un mouvement saisonnier stable. Ce test est appliqué à des séries formées de composantes saisonnières estimées et d'aléas (résidus) qui sont très susceptibles d'être autocorrélés, ce qui va à l'encontre de l'hypothèse fondamentale du test  $F$ . Les producteurs de données désaisonnalisées connaissent depuis longtemps cette lacune et se servent rarement de la valeur théorique de la statistique  $F$  comme critère pour la désaisonnalisation. Ils préfèrent utiliser des règles empiriques du genre "  $F$  égal ou supérieur à 7 ". Dans cet article, nous présentons un test exact qui tient compte des résidus autocorrélés qui suivent un processus MMS (à moyennes mobiles saisonnier) du type  $(0, q)(0, \bar{Q})_s$ . Nous comparons ensuite les résultats de ce test, qui est une version modifiée du test  $F$ , avec ceux du test d'analyse de variance de la X-11-ARMMI pour un grand nombre de séries socio-économiques canadiennes.

MOTS CLÉS: Analyse de variance ordinaire; résidus autocorrélés; mouvement saisonnier.

## 1. INTRODUCTION

Dans l'analyse des séries économiques et sociales, on décompose normalement la série de données observées en quatre facteurs non observés: la tendance, le cycle, les variations saisonnières et les aléas.

Les séries socio-économiques sont souvent présentées sous forme de désaisonnalisées de sorte qu'il soit plus facile d'analyser la tendance conjoncturelle et que l'on puisse évaluer les conditions socio-économiques courantes. Il existe plusieurs méthodes de désaisonnalisation qui permettent d'estimer la composante saisonnière d'une série mais la variante X-11 du programme *Census Method II* (Shiskin, Young et Musgrave 1967) et la méthode X-11-ARMMI (Dagum 1980) sont les plus utilisées. Pour déterminer la présence d'un mouvement saisonnier stable dans une série chronologique, la X-11-ARMMI et la variante X-11 utilisent toutes deux les résultats du test  $F$  appliqué habituellement dans une analyse de variance à un critère de classification pour les variations saisonnières mensuelles et les résidus. Or, les résidus étudiés dans cette analyse de variance sont souvent autocorrélés, de sorte que le seuil de signification théorique du test  $F$  peut ne pas être bon. Conscients de cette lacune, les producteurs de données désaisonnalisées ne se fondent pas sur ce seuil pour vérifier la présence d'un mouvement saisonnier stable mais appliquent plutôt une règle basée sur des connaissances empiriques (voir, par exemple, Shiskin et Plewes 1978). De fait, le test X-11-ARMMI de la présence d'un "mouvement saisonnier identifiable" suppose implicitement que le valeur  $F$  sera égale ou supérieure à 7 s'il n'existe pas de mouvement saisonnier évolutif.

Les tests de la présence d'un mouvement saisonnier stable (comme ceux visant à vérifier l'existence de variations saisonnières annuelles) peuvent être assimilés aux tests qui servent à

<sup>1</sup> Brajendra C. Sutradhar, Département de mathématique et de statistique, Memorial University of Newfoundland, St.-Jean, Terre-Neuve, A1C 5S7; Estela Bee Dagum, Division des séries chronologiques - Recherche et analyse, Statistique Canada, Ottawa, Ontario, K1A 0T6; Binyam Solomon, Direction de l'analyse économique et sociale, Quartier général de la Défense nationale, Ottawa, Ontario, K1A 0K2.





REMERCIEMENTS

Les auteurs désirent remercier le professeur Fay Cook, le professeur Christopher Jencks, le professeur Dan Lewis et le professeur Dennis Rosenbaum, qui leur ont donné accès aux ensembles de données utilisés pour les analyses secondaires dont fait état cet article. Les auteurs souhaitent aussi remercier le professeur Peter V. Miller pour les précieux commentaires qu'il a faits sur une version antérieure du texte de cet article.

BIBLIOGRAPHIE

BURGESS, R.D. (1989). Major issues and implications of tracing survey respondents. Dans *Panel Surveys*, (Eds. D. Kasprzyk, G. Duncan, G. Kalton et M.P. Singh). New York: John Wiley & Sons.

CRANO, W.D., et BREWER, M.B. (1973). *Principles of Research in Social Psychology*. New York: McGraw-Hill.

CRIDER, D. M., WILLETS, F.K., et BEALER, R.C. (1971). Tracking respondents in longitudinal surveys. *Public Opinion Quarterly*, 35, 613-620.

DROEGE, R.C., et CRAMBERT, A.C. (1965). Follow-up techniques in a large-scale test validation study. *Journal of Applied Psychology*, 49, 253-256.

FREEMAN, D.S., THORTON, A., et CAMBURN, D. (1980). Maintaining response rates in longitudinal studies. *Sociological Methods & Research*, 9, 87-98.

HAUSMAN, J.A., et WISE, D.A. (1979). Attrition bias in experimental and panel data: the Gary income maintenance experiment. *Econometrica*, 47, 455-473.

LAVRAKAS, P.J. (1987). *Telephone Survey Methods: Sampling, Selection and Supervision*. Newbury Park, CA: Sage.

LEHNEN, R.G., et KOCH, G.G. (1974). Analyzing panel data with uncontrolled attrition. *Public Opinion Quarterly*, 38, 40-56.

LEPKOWSKI, J.M. (1989). Treatment of wave nonresponse in panel surveys. Dans *Panel Surveys*, (Eds. D. Kasprzyk, G. Duncan, G. Kalton et M.P. Singh). New York: John Wiley & Sons.

MCALLISTER, R.J., GOE, S.J., et BUTLER, E.W. (1973). Tracking respondents in longitudinal surveys: some preliminary considerations. *Public Opinion Quarterly*, 37, 413-416.

SOBOL, M.G. (1959). Panel mortality and panel bias. *Journal of the American Statistical Association*, 54, 52-68.

WINER, R.S. (1983). Attrition bias in econometric models estimated with panel data. *Journal of Marketing Research*, 177-186.

À l'ère du micro-ordinateur, il serait tout à fait possible de donner à l'intervieweur un "profil" de chaque répondant au cycle 1 pour qu'il puisse se familiariser avec la personne à réinterviewer. Il faudrait se garder de créer chez l'intervieweur des attentes qui pourraient biaiser les réponses des enquêtes au cycle 2. Nous ne voulons pas dire que l'intervieweur utiliserait nécessairement ces renseignements mot à mot pour identifier le répondant du cycle 1 au moment d'entrer de nouveau en contact avec lui; nous croyons que le nom, le sexe et l'âge suffisent pour cela. Mais, si l'intervieweur a une idée plus précise de "qui" est le répondant, il peut y avoir dans son comportement verbal un changement subtil susceptible de faire augmenter son taux de réussite. Cette suggestion doit subir le test de l'expérience avant qu'on puisse la recommander en toute confiance, mais s'il s'avérait qu'elle est efficace et n'introduit pas de biais dans les données, elle serait relativement facile à appliquer.

De même, l'entrée en matière que l'intervieweur lit au répondant au moment du cycle 2 pourrait être adaptée aux groupes démographiques qui paraissent le plus susceptibles de refuser de répondre au cycle 2; nous pensons ici aux personnes âgées, à celles qui ont un niveau d'ins-truction peu élevé ou dont le revenu est relativement faible et plus spécialement à celles que les intervieweurs du cycle 1 auraient jugées peu intéressées ou disposées à coopérer. Ici encore, on pourrait programmer un ordinateur qui pourrait générer des entrées en matière à partir des données du cycle 1 relatives à certains répondants.

Les entrées en matière doivent contenir des éléments qui incitent ces personnes à participer au cycle 2, car souvent ce sont elles qui ont le moins de motivation intrinsèque pour participer à une enquête. Au moment de préparer les cycles subséquents, les concepteurs de l'enquête doivent se demander "pour quelles raisons" ces personnes pourraient vouloir offrir leur col-laboration et incorporer ces raisons dans les entrées en matière qui leur sont destinées. Cel-les-ci pourraient être assez longues et même contenir des questions destinées à établir un rapport avec l'enquête. Peut-être même serait-il possible – sans introduire de biais dans les réponses du cycle 2 – de donner à la personne dont on souhaite obtenir des réponses quelques rensei-gnements sur les résultats du cycle 1. Peut-être le répondant serait-il alors plus enclin à consi-dérer comme un "échange" la tentative de reprise de contact.

Quoi qu'il en soit, l'ordinateur pourrait générer ces entrées en matière spéciales, qui seraient utilisées seulement pour les répondants auxquels elles sont destinées. Encore une fois, aucune donnée expérimentale ne peut confirmer l'efficacité de cette idée, mais nous croyons qu'elle mérite d'être étudiée.

## 5. CONCLUSION

Les observations que nous avons faites ici donnent à penser que dans les enquêtes par panel par composition aléatoire où les noms des répondants ne sont pas connus, la perte d'effectif n'est pas suffisante pour invalider ou rendre inutilisable cette technique d'enquête. Comme celle-ci ne provoque pas chez le répondant une réaction de crainte d'être évalué, elle paraît être l'approche à privilégier pour les enquêtes par panel à deux cycles par composition aléatoire, dans le cas où le chercheur a des raisons *a priori* de ne pas vouloir que le répondant sache qu'il sera réinterviewé. Nos observations devraient aussi être encourageantes pour ceux qui envisa-gent de convertir en enquête par panel une enquête transversale par composition aléatoire. Nous espérons que cet article essentiellement descriptif encouragera d'autres spécialistes des métho-des d'enquête à faire, dans des conditions mieux contrôlées, des études sur la nature et l'import-tance de la perte d'effectifs dans les enquêtes par panel par composition aléatoire et à com-muniquer les résultats de ces études; ainsi, les chercheurs pourront peut-être un jour appliquer avec plus de confiance des stratégies de réduction des pertes d'effectifs. Nous pensons que cette recherche devrait être guidée par un fait d'expérience, qui est que la réduction de la perte d'effec-tifs paraît le plus efficace dans les enquêtes bien organisées où chaque répondant est considéré comme la personne individuelle qu'il est en fait.



à l'aide des annuaires téléphoniques ou en composant les nouveaux numéros donnés par les messages enregistrés des compagnies de téléphonie ou même en appelant d'anciens voisins pour obtenir un numéro de renvoi quand l'adresse d'un répondant est connue et qu'on se sert d'un annuaire par numéros. Mais si l'on n'a pas l'intention de retracer les répondants au cycle 2, pourquoi leur demander leur nom au cycle 1 ?

Sans doute les intervieweurs préféreraient-ils avoir le nom au complet: la plupart se sentent plus à l'aise s'ils doivent demander à parler à "Pierre" ou à "Pierre Tremblay" plutôt qu'à "un homme au milieu de la cinquantaine". Pourtant, la faible différence entre les taux de perte d'effectifs dans les deux études qui nous occupent ici ne démontre pas avec netteté qu'il est avantageux de demander leurs noms aux répondants, et cela même en tenant compte du fait que le décalage entre les cycles est plus long de quatre mois dans l'étude 2, dans laquelle des descripteurs de nom avaient été recueillis au cycle 1. Nous reconnaissons qu'une des limites regrettables de notre article est que d'autres différences entre ces deux enquêtes par panel par composition aléatoire ont pu influencer sur l'écart observé entre les taux de perte d'effectifs. Par exemple, le nombre de rappels effectués au cycle 2 était beaucoup plus élevé dans l'étude 2 que dans l'étude 1 (plus de vingt contre huit). La question restera donc sans réponse tant que nous ne seront pas effectuées des expériences mieux contrôlées.

Dans l'état actuel des connaissances, nous croyons que c'est à chaque chercheur qui utilise un panel aux fins d'une enquête par composition aléatoire qu'il appartient de comparer, d'une part, le risque de biaiser les données sur le phénomène étudié, s'il avertit les répondants qu'ils seront de nouveau "mesurés" plus tard (effet de la "réaction" du répondant) et, d'autre part, la possibilité d'obtenir un taux de perte d'effectifs un peu plus faible en demandant leurs noms aux répondants au moment de l'interview du cycle 1.

**Considérations sur le moyen de réduire au minimum les effets de la perte d'effectifs.** On peut examiner un certain nombre de moyens pouvant servir à réduire au minimum les effets de la perte d'effectifs dans une enquête par panel par composition aléatoire.

Sobol (1959) a suggéré la possibilité d'un *suréchantillonnage*, au cycle 1, des répondants qui ont les caractéristiques les rendant le plus susceptibles d'être perdus aux cycles suivants. À première vue, cette idée peut paraître attrayante: en effet, si l'on sait quels répondants sont le plus susceptibles de n'être pas réinterviewés, on peut projeter de suréchantillonner ces groupes au cycle 1. Comme le travail de Sobol l'a montré et comme le confirment les résultats des deux études considérées ici, il serait possible d'estimer quels types de personnes il faudrait suréchantillonner au cycle 1: par exemple, les jeunes adultes et les adultes âgés. Le suréchantillonnage pourrait se faire à l'aide d'une technique de sélection des répondants qui serait appliquée vers la fin du cycle 1 (mais qui augmenterait le coût global de ce cycle 1 de l'enquête). Même s'il est possible, un tel suréchantillonnage est-il souhaitable? Poser cette question revient, en fin de compte, à se demander si le panel ainsi obtenu est plus qu'une réflexion *superficielle* de la population étudiée. En d'autres termes, suffit-il de se préoccuper uniquement d'avoir le nombre voulu (c'est-à-dire la bonne proportion) de personnes âgées au dernier cycle d'une enquête par panel, ou bien ne faut-il pas aussi se demander si la "composition" du groupe de ces personnes âgées est celle qui convient?

Il s'agit là d'une question empirique à laquelle les études actuelles ne peuvent pas répondre. Il est clair que d'autres recherches devront être faites pour que les spécialistes des enquêtes puissent déterminer avec plus de certitude s'il est préférable de suréchantillonner au cycle 1 ou de "compenser" la perte d'effectifs par des corrections statistiques des données recueillies aux cycles subséquents de l'enquête par panel.

Un autre aspect du problème de la perte d'effectifs est lié aux efforts visant à réduire au minimum la perte de répondants que les intervieweurs arrivent à joindre au cycle 2 mais qui refusent de se laisser réinterviewer ou ne sont "jamais à la maison" pour répondre aux questions du cycle 2. Ces cas représentaient 29% de la perte d'effectifs dans l'étude 1 et 34% dans l'étude 2. Hors les techniques d'interview habituelles et la répétition des appels de relance, quels moyens peut-on donner aux intervieweurs pour qu'ils parviennent mieux à réduire ces pertes?

Pour évaluer correctement cette faible différence entre les taux de perte d'effectifs ( $\chi^2 (1) = 3,51, p > .10$ ), il faut tenir compte des différences de contexte suivantes entre les deux études. Dans l'étude 1, on ne disait pas explicitement aux répondants du cycle 1 qu'on les rappellerait un an plus tard et on ne leur demandait donc pas leur nom. Dans l'étude 2 au contraire, on disait aux répondants qu'on les rappellerait dans un certain temps. Etant donné la nature particulière de la recherche dans l'étude 1, on n'a pas tenté de retracer les répondants du cycle 2 qui avaient déménagé ou changé de numéro de téléphone, alors que cela fut fait dans l'étude 2, bien qu'avec assez peu de résultats. Dans l'étude 1, il y avait aussi une version espagnole du questionnaire, tandis que dans l'étude 2 les hispanophones ne parlaient pas l'anglais n'étaient pas interviewés.

Dans les deux études, la très grande majorité des répondants perdus pour l'enquête étaient des personnes qu'on ne pouvait joindre à leur numéro de téléphone du cycle 1 parce que ce numéro était celui d'un nouveau logement, parce que le répondant avait déménagé ou parce que le numéro n'était plus en service.

Dans l'ensemble, les résultats de ces deux enquêtes téléphoniques s'accordent assez bien avec ceux de précédentes enquêtes effectuées au moyen d'interviews en personne (voir, par exemple, Sobol 1959), du point de vue des caractéristiques des personnes les plus susceptibles de n'être pas réinterviewées au moment d'un cycle subséquent d'une enquête par panel. Dans les deux études, les jeunes adultes et les adultes âgés, les personnes de couleur, les personnes ayant un niveau d'instruction peu élevé et les personnes à faible revenu étaient moins susceptibles d'être réinterviewées que les autres sous-groupes démographiques de la même catégorie.

## 4.2 Conséquences

Comme les enquêtes par composition aléatoire offrent un bilan coûts-avantages intéressant et qui vient s'ajouter aux avantages que comportent les enquêtes par panel du point de vue des possibilités d'analyse, il n'est pas inutile d'examiner les options qui pourraient améliorer la représentativité du panel final dans les enquêtes où l'échantillon du cycle 1 est obtenu par composition aléatoire. Mais avant d'entrer dans ces considérations, il convient de nous arrêter plus longtemps au problème de savoir s'il faut ou non demander les noms des répondants dans les enquêtes téléphoniques.

**Doit-on demander le nom du répondant au cycle 1?** Comme nous l'avons dit, la question pourrait se poser en ces termes: ou bien augmenter la probabilité d'atteindre et, donc, de réinterviewer le répondant au cycle 2, ou bien risquer de susciter chez lui une réaction de crainte d'être évalué (Crano et Brewer 1973), qui aurait pour effet de baisser les données recueillies au cycle 2. Mais il y a plus que cette alternative à considérer.

La confidentialité des données et la possibilité pour le répondant de donner son autorisation en connaissance de cause sont des questions qui entrent également en ligne de compte. La pratique courante parmi les organismes qui effectuent des enquêtes statistiques pour des universitaires est de ne jamais communiquer les numéros de téléphone à quiconque, sauf peut-être au promoteur de l'enquête et seulement si ce dernier projette d'effectuer une enquête par panel ou des interviews de suivi auprès de répondants qui l'ont explicitement autorisé. Le principe derrière cette pratique est que l'assurance donnée aux répondants du cycle 1 quant au caractère confidentiel de leurs réponses est respectée si on les rappelle dans le cadre de la même enquête. Le fait qu'il y ait si peu de répondants du cycle 1 qui refusent de participer au cycle 2 et qu'il soit possible de prédire, à partir de leurs caractéristiques démographiques, quelles personnes sont les plus susceptibles de refuser de répondre au moment du cycle 2 donne beau-coup de poids à la conclusion selon laquelle une seconde interview téléphonique dont le répondant n'aurait pas été informé au moment du cycle 1 ne pose pas de difficulté.

Quand le promoteur d'une enquête peut supporter le coût supplémentaire qu'entraîne le fait de retracer les répondants qui ont déménagé, il paraît logique de noter le nom complet des répondants au moment du cycle 1 puisqu'il est alors possible de retracer ceux qui ont déménagé.



c'était possible; mais cet effort n'a pas donné beaucoup de résultats puisque, dans la plupart des cas, on n'avait pas le nom complet (prénom et nom de famille) du répondant. Comme pour le cycle 1, au moins vingt appels étaient effectués pour les répondants les plus difficiles à joindre. Plus de 80% des répondants avaient donné un descripteur de nom lors du cycle 1. Les intervieweurs utilisaient ce renseignement comme suit:

Bonjour! Suis-je bien au \_\_\_\_\_? Je m'appelle \_\_\_\_\_, et je vous téléphone de l'Université Northwestern. À ce même numéro il y a environ 16 mois – soit vers la fin de 1983 – nous avons interviewé (un homme/une femme) du nom de \_\_\_\_\_. Pourrais-je lui parler?

Comme dans l'étude 1, l'intervieweur vérifiait d'abord le numéro, puis donnait son nom. Le troisième espace blanc contenait le descripteur de nom recueilli au moment du cycle 1. Pour les répondants qui n'avaient pas donné leur nom au cycle 1, on utilisait la même méthode que dans l'étude 1 pour les demander au téléphone (c'est-à-dire qu'on mentionnait le sexe et l'âge ou le sexe seulement).

### 3.2 Résultats

**Importance numérique de la perte d'effectifs.** Comme le montre le tableau 1, près de 60% des répondants du cycle 1 ont été réinterviewés (57.4%). Dans l'ensemble, les fréquences relatives des causes des pertes d'effectifs dans l'étude 2 étaient très semblables à celles de l'étude 1. Parmi les 425 répondants perdus pour l'enquête, la plus grande partie (près de 40%) était considérée des cas où le numéro de téléphone correspondait à un nouveau ménage ou à un logement que le répondant du cycle 1 avait quitté sans laisser de nouveau numéro où le joindre. Le deuxième cas en importance – presque un quart des pertes – était celui des personnes dont le numéro de téléphone au cycle 1 n'était plus en service. Suivait la catégorie des répondants qui d'une manière ou d'une autre refusaient de répondre. La quatrième catégorie était celle des numéros du cycle 1 où il n'y avait jamais eu de réponse au cycle 2.

**Nature de la perte d'effectifs.** Comme le montre le tableau 1, plusieurs des variables se rapportant au groupe des répondants du cycle 2 de l'étude 2 qui ont pu être interviewées différemment de celles qui se rapportent au groupe de ceux qui ont été perdus pour l'enquête, avec des variations très semblables à celles qui ont été observées pour l'étude 1. En ce qui concerne l'âge, les adultes qui avaient moins de 34 ans ou plus de 64 ans au cycle 1 étaient moins susceptibles d'être réinterviewés que les autres. Les personnes d'origine asiatique et les Blancs. En ce qui concerne le niveau de scolarité, les répondants qui avaient peu d'années de scolarité étaient moins susceptibles d'être réinterviewés que ceux qui en avaient davantage. À peine 50% des répondants qui avaient indiqué un revenu du ménage inférieur à \$12,000 lors du cycle 1 ont pu être réinterviewés, alors que ceux qui avaient un revenu de plus de \$24,000 ont pu être réinterviewés dans une proportion de 64%. Les répondants divorcés ont eu le meilleur taux de suivi (68%), tandis que ceux qui avaient dit être séparés au moment du cycle 1 étaient les moins susceptibles d'être réinterviewés (43%).

## 4. ANALYSE

### 4.1 Sommaire des résultats

Les deux études indépendantes analysées ici sont des enquêtes par composition aléatoire à deux cycles, avec un décalage, l'une de douze et l'autre de seize mois, entre les cycles. Dans l'étude 1, où les noms des répondants n'étaient pas connus et ne pouvaient donc pas servir au cycle 2, la perte d'effectifs avait été de 47.1%. Dans l'étude 2, où l'on avait des descripteurs de nom pour 83% des répondants du cycle 1, la perte d'effectifs avait été légèrement inférieure, soit de 42.6%.



Tableau 2

Caractéristiques des répondants à la réinterview du cycle 2 pour l'étude 1 (noms connus) et l'étude 2 (noms inconnus)

Pourcentage de personnes réinterviewées

Caractéristiques des répondants

Étude 1

Étude 2

Sexe:

Femmes

Hommes

55

49

60

Age:

< 30 ans

30-39 ans

40-59 ans

> 59 ans

< 34 ans

35-49 ans

50-64 ans

> 64 ans

Race:

Asiatique

Noir

Hispanique

Blanc

68\*\*\*

49

44

58

Niveau d'instruction:

Pas de diplôme d'études secondaires

Diplôme d'études secondaires

Études postsecondaires partielles

Diplôme d'études postsecondaires

Études supérieures

Revenu du ménage:

< \$10,000

\$10,000-\$19,999

\$20,000-\$29,999

\$30,000 ou plus

< \$12,000

\$12,000-\$17,999

\$18,000-\$23,999

\$24,000 ou plus

État matrimonial:

Marié

Divorcé

Séparé

Célibataire

Veuf/veuve

Jamais marié

Mode d'occupation du logement:

Propriétaire

Locataire

Nombre d'années de résidence dans le quartier:

< 3 ans

3-9 ans

10 ans ou plus

Déménagement dans les deux prochaines années:

Absolument certain

Probable

Improbable

Absolument exclu

37\*\*\*

47

57

60

Remarque: Des tests de signification du Khi carré ont été utilisés.

\*\*\*  $p < .001$

\*\*  $p < .01$

\*  $p < .05$

## 2.2 Résultats

À cause d'une erreur dans le traitement des questionnaires et des registres d'appels du cycle 1, des numéros d'identification erronés ont été attribués à dix-sept répondants du cycle 1 par le personnel du promoteur de l'enquête. Aux fins de cet article, notre analyse n'a pas tenu compte de ces répondants parce que nous ne pouvions pas associer avec certitude les réponses du cycle 2 aux données correspondantes du cycle 1. Les analyses qui suivent portent donc sur les 797 répondants pour lesquels la correspondance des données du cycle 1 à celles du cycle 2 est certaine.

**Importance numérique de la perte d'effectifs.** Comme le montre le tableau 1, environ la moitié (53%) de l'échantillon du cycle 1 a pu être réinterviewé. Parmi les 375 répondants "perdus" pour l'enquête, la proportion la plus élevée était composée de numéros de téléphone qui correspondaient à de nouveaux ménages ou à des logements du cycle 1 que n'habitaient plus les répondants; ces cas représentaient environ 40% des pertes. Le deuxième cas en importance était celui des personnes dont le numéro de téléphone au cycle 1 n'était plus en service; ceux-là représentaient environ un quart des pertes. Venaient ensuite les répondants qui avaient refusé d'une façon ou d'une autre de répondre. Au quatrième rang venaient les personnes qui, en huit tentatives, n'avaient jamais été à domicile au moment où quelqu'un prenait l'appel du cycle 2. (Pour chacune des 33 personnes dans ce cas, un autre membre de leur ménage a indiqué qu'il s'agissait bien du répondant du cycle 1).

**Nature de la perte d'effectifs.** Comme on peut le voir au tableau 2, plusieurs des variables se rapportant au groupe des répondants du cycle 1 qui ont pu être réinterviewées diffèrent sensiblement de celles qui se rapportent au groupe des personnes qui ont été perdues pour l'enquête. En ce qui concerne l'âge, on a pu interviewer une deuxième fois seulement 42% des adultes de moins de 30 ans, alors que 60% des adultes de 40 à 59 ans ont répondu aux questions du cycle 2. Les Noirs étaient beaucoup moins susceptibles que les Blancs d'être réinterviewés. Quant au revenu du ménage, les répondants qui au cycle 1 avaient déclaré un revenu de moins de \$10,000 n'ont été réinterviewés que dans 44% des cas, alors que 63% de ceux qui avaient un revenu de plus de \$20,000 ont pu être réinterviewés. On a réinterviewé plus de répondants mariés (57%) que de répondants non mariés (49%). Les répondants propriétaires ont répondu au deuxième questionnaire dans une proportion de 62%, comparativement à 47% des répondants locataires. Plus un répondant avait vécu longtemps dans le quartier – et plus il avait, au cycle 1, déclaré improbable qu'il puisse déménager – plus il était susceptible d'être réinterviewé.

## 3. ETUDE 2

### 3.1 Méthodologie

En novembre et décembre de 1983, une enquête par composition aléatoire (à un degré) a été menée par le Northwestern University Survey Laboratory pour des professeurs qui voulaient étudier le bien-être économique des familles à Chicago. (L'interview à l'aide du questionnaire prenait 20 minutes en moyenne.) Environ 3,900 numéros de téléphone ont été composés et 997 personnes ont été interviewées. Pour chaque unité d'habitation contactée, on sélectionnait systématiquement un chef de ménage (homme ou femme) comme répondant désigné. On rappelait jusqu'à vingt fois pour augmenter la probabilité d'obtenir une interview auprès des répondants difficiles à joindre. En tout, on a atteint 1,659 ménages admissibles; quant aux personnes admissibles qui n'ont pas été interviewées, elles n'étaient pas disponibles au moment des appels ou ont refusé de participer.

Seize mois plus tard, au printemps de 1985, on a de nouveau composé les 997 numéros de téléphone pour recueillir les données du cycle 2. Contrairement à ce qu'on avait fait dans l'étude 1, où l'on ne tentait pas de retracer les répondants qui avaient déménagé ou changé de numéro de téléphone, un effort a été fait dans l'étude 2 pour retracer les répondants quand

Bonjour! Suis-je bien au \_\_\_\_\_? Je m'appelle \_\_\_\_\_, et je vous téléphone de l'Université Northwestern. À ce même numéro, il y a environ un an – soit en février 1983 – nous avons interviewé une( ) \_\_\_\_\_.

Pourrais-je lui parler?

L'intervieweur vérifiait d'abord le numéro, puis donnait son nom. Le troisième *espace blanc* contenait les renseignements démographiques recueillis lors du cycle 1 (sexe et âge) sur cha-que répondant: par exemple, "femme au milieu de la trentaine" ou "homme d'un peu plus de 70 ans". Pour les quelques répondants qui n'avaient pas donné leur date de naissance lors du cycle 1, le troisième espace blanc contenait seulement la mention "homme" ou "femme".

Une fois que l'intervieweur savait qu'il parlait au répondant du cycle 1, il poursuivait en donnant l'explication suivante, avant de commencer l'interview:

Les renseignements que vous nous avez donnés l'an dernier nous ont beaucoup aidés à comprendre les préoccupations de résidents de Chicago comme vous. Nous vous rappelons maintenant pour avoir des renseignements sur ce qu'a été la qualité de la vie dans les différents quartiers de Chicago depuis un an.

Le but de cette entrée en matière était de mieux disposer le répondant à coopérer avec l'intervieweur du cycle 2 en lui rappelant la collaboration qu'il avait offerte lors du cycle 1.

Vu les objectifs de ce projet d'évaluation, les répondants qui avaient déménagé ou dont le numéro de téléphone avait changé n'étaient pas réinterviewés pour le cycle 2: il fallait en effet interviewer seulement des personnes qui habitaient à la même adresse que l'année précédente, étant donné que plusieurs questions avaient trait à la perception d'un changement dans le voi-sinage depuis 1983.

Tableau 1

Résultat de la tentative de réinterview des répondants au cycle 2 de l'étude 1 (noms inconnus) et de l'étude 2 (noms connus)					
Résultat au cycle 2			Résultat au cycle 2		
Étude 1 – Noms inconnus			Étude 2 – Noms connus		
Fréquence absolue	Fréquence relative	%	Fréquence absolue	Fréquence relative	%
Aucun contact établi au cycle 2					
Téléphone hors d'usage					
ou débranché					
95	11.9	2.1	94	45	4.5
17					
Jamais de réponse					
Contact établi					
Interview a eu lieu					
422	52.9		572		57.4
165	20.7		163		16.4
33	4.1		30		3.0
Répondant jamais disponible					
Réfusal total ou partiel du					
répondant					
37	4.7		57		5.7
21	2.6		13		1.3
3	0.4		13		1.3
4	0.6		10		1.0
Autre					
797	100.0		997		100.0
Total					



Comment peut-on atteindre de nouveau un répondant quand on ne lui a pas demandé son nom la première fois? C'est là un problème auquel le premier auteur a eu à faire face en 1979 lorsqu'il essayait de déterminer s'il était possible de transformer en enquête par panel une enquête transversale de 1977. Comme il n'y avait pas de travaux publiés qui auraient pu l'aider dans sa tâche, il a procédé à une expérience pilote: il a été possible, en mentionnant le sexe et l'âge, de réinterviewer 50% des répondants de 1977.

Les résultats de cette expérience pilote étaient assez encourageants pour que la méthode soit réutilisée dans la première des deux études qui font l'objet de cet article. Dans l'étude 1, les intervieweurs composaient donc le numéro de téléphone qui avait servi à l'interview du cycle 1, vérifiaient ce numéro lorsqu'ils obtenaient une réponse et informaient l'interlocuteur qu'envoyé un an auparavant, au même numéro de téléphone, quelqu'un avait répondu à une enquête téléphonique. L'intervieweur identifiait le répondant du cycle 1 en précisant le sexe "homme" ou "femme" et l'âge (par exemple "début de la vingtaine" ou "fin de la soixantaine") de la personne à qui il voulait parler.

Dans l'étude 2, on possédait un descripteur de nom pour plus de 80% des répondants du cycle 1. Les intervieweurs demandaient ces répondants au téléphone en utilisant le descripteur de nom, après avoir vérifié le numéro de téléphone. Quant aux répondants pour lesquels on n'avait pas de descripteur de nom, les intervieweurs les demandaient à l'aide de descripteurs démographiques, comme dans l'étude 1.

En présentant les résultats de cette étude, nous souhaitons seulement donner une première idée de l'importance numérique et de la nature de la perte d'effectifs dans les enquêtes par panel par composition aléatoire. Bien qu'on ne puisse les étendre à un échantillon national obtenu par composition aléatoire, ces résultats sont parlants. Comme l'échantillonnage par composition aléatoire est très utilisé, nous croyons qu'il importe de construire une base de connaissances sur la perte probable d'effectifs dans des études par panel où l'échantillonnage du cycle 1 est fait par composition aléatoire, particulièrement lorsque les chercheurs n'ont pas de descripteur de nom pour les répondants du cycle 1. Ainsi, il sera plus facile d'examiner des stratégies conçues pour réduire l'importance et les effets de cette perte d'effectifs.

## 2. ETUDE 1

### 2.1 Méthodologie

En février 1983, le laboratoire d'enquêtes de l'Université Northwestern a mené une enquête par composition aléatoire (à un degré) portant sur toute la ville, afin de recueillir des données de base pour des professeurs qui voulaient évaluer un ensemble de programmes locaux de prévention du crime dans des certains quartiers de Chicago. (L'interview à l'aide du questionnaire prenait 20 minutes en moyenne.) Environ 2,800 numéros de téléphone ont été composés et 814 personnes interviewées. Pour chaque unité de logement contactée, on sélectionnait systématiquement un chef de ménage (homme ou femme) comme répondant désigné (Lavrakas 1987; p. 99-100). Chaque fois que cela s'avérait nécessaire, un questionnaire en langue espagnole était utilisé par un intervieweur bilingue. On rappelait jusqu'à sept fois les répondants difficiles à joindre. Parmi les numéros composés, 1,247 correspondaient à des ménages admissibles (définis par les promoteurs de l'enquête comme des ménages anglophones ou hispanophones) comptant au moins un adulte âgé de 19 ans ou plus; quant aux ménages admissibles pour lesquels il n'y a pas eu d'interview, ou bien la personne qui aurait pu répondre n'était pas disponible au moment de l'appel ou bien elle a refusé de participer.

Un an plus tard, en février 1984, on a de nouveau composé les numéros de téléphone du cycle 1 pour recueillir des données "post enquête" pour le projet d'évaluation. Si quelqu'un répondait au téléphone après huit tentatives ou moins (faites sur plusieurs jours et à des heures différentes), l'intervieweur lisait le message de présentation suivant:

Sobol signale qu'en général, "à cause des variations qui s'annulent l'une l'autre, la structure démographique, après cinq rondes d'entrevues, demeurerait très semblable à celle de l'échantillon original" (p. 52). Il s'est pourtant produit quelques variations significatives, la perte d'effectifs ayant été disproportionnée parmi les locataires, les ménages à faible revenu, les résidents des grandes régions métropolitaines, les jeunes adultes (moins de 25 ans), les personnes âgées (plus de 64 ans) et les personnes que n'intéressaient pas le sujet de l'enquête. Winer (1983) a fait état de résultats d'études non publiées qui confirment généralement les résultats obtenus par Sobol.

Il importe de remarquer que dans chacune de ces études, les intervieweurs connaissaient le nom complet de chaque répondant du cycle 1. En fait, dans leur article sur les techniques acceptées de réduire au minimum la perte d'effectifs dans une enquête par panel, McAllister *et al.* (1973) ont insisté sur l'importance de recueillir, à la fin de l'interview, des renseignements détaillés sur les coordonnées futures du répondant, y compris "les noms et les adresses complètes d'amis ou de parents . . . du répondant." (p. 416).

Bien qu'on puisse faire valoir que la perte d'effectifs est un problème suffisamment grave pour qu'au cycle 1 les chercheurs jugent nécessaire de demander au répondant son nom et d'autres renseignements permettant de l'identifier, cette approche peut elle-même causer des problèmes. Dans les cas où l'on demande son nom au répondant au moment de l'entrevue du cycle 1, on lui explique parfois que ce renseignement est important parce qu'il sera ou pourra être rappelé après un certain temps pour qu'on puisse déterminer s'il s'est produit des changements. Cela soulève la question de la réaction de l'enquête, c'est-à-dire de sa "crainte d'être évaluée". (Voir Crano et Brewer 1973.) Certains auteurs étudient de façon explicite l'alternativité que détermine la nécessité de choisir entre la perte d'effectifs et les effets de cette crainte (voir par exemple Sobol 1959), mais la plupart supposent implicitement que la crainte de l'évaluation est un problème moins grave que la perte d'effectifs.

Tous les travaux mentionnés ci-dessus ont été effectués à l'aide d'interviews en personne. Mais quelle est la perte d'effectifs dans les enquêtes téléphoniques, y compris dans les cas où l'on ne demande pas leurs noms aux répondants du cycle 1? En particulier, à quoi peut s'attendre un chercheur qui décide *a priori* de faire une enquête téléphonique par panel et demande donc des descripteurs de nom aux répondants, par opposition à un autre chercheur qui ne demande pas de descripteurs de nom parce qu'il a choisi de ne pas le faire ou décide après coup de transformer une enquête téléphonique transversale en une enquête par panel, après la fin des interviews du cycle 1?

Le présent article tente de définir une première approche de ces problèmes en présentant, relativement à l'importance numérique et à la nature de la perte d'effectifs, les résultats de deux enquêtes par panel par composition aléatoire effectuées dans la ville de Chicago, avec un décalage d'environ un an entre les cycles dans chacun des cas. Notons que ces deux enquêtes ont été menées de façon indépendante et n'étaient pas conçues pour servir à une expérience sur la perte d'effectifs dans les enquêtes par composition aléatoire. Ainsi, il y a des différences du point de vue du centre d'intérêt et des modalités d'exécution entre ces deux enquêtes, qui s'ajoutent au fait qu'on avait un descripteur de nom pour la plupart des répondants dans l'étude 2 mais non dans l'étude 1. Nous reconnaissons explicitement que ces différences de centre d'intérêt et de modalités d'exécution limitent les conclusions que l'on peut tirer de la comparaison des deux études.

Dans les deux études, l'échantillonnage des répondants du cycle 1 s'est fait par composition aléatoire à un degré (dite simple). Dans l'étude 1, on ne demandait pas son nom au répondant du cycle 1 et on ne lui disait pas qu'on entrerait de nouveau en contact avec lui. Dans l'étude 2, des descripteurs de nom étaient recueillis à la fin de l'interview du cycle 1, et ces descripteurs étaient utilisés pour entrer en contact avec les répondants au moment du cycle 2. Au lieu de donner son nom au complet, le répondant ne donnait le plus souvent que son prénom ou un autre descripteur, par exemple un surnom ou des initiales. (L'intervieweur n'insistait pas pour obtenir le nom complet afin de ne pas alimenter les sentiments de crainte que pouvait avoir un répondant réticent.)



## Perte d'effectifs dans un panel obtenu par composition aléatoire dans deux enquêtes locales

PAUL J. LAVRAKAS, RICHARD A. SETTERSTEN, Jr.  
et RICHARD A. MAIER, Jr.<sup>1</sup>

### RÉSUMÉ

Cet article compare, du point de vue de la nature et de l'importance numérique de la perte d'effectifs, deux enquêtes par panel par composition aléatoire (CA) menées séparément dans la ville de Chicago (c'est-à-dire que les enquêtes étaient des études indépendantes et n'étaient pas menées dans le cadre d'une même expérience) et ayant l'autre un décalage d'environ un an entre les cycles. Pour chaque enquête, l'échantillonnage au cycle 1 se faisait par composition aléatoire à un degré (échantillonnage aléatoire). Dans l'étude 1, on ne demandait pas son nom au répondant, de sorte qu'au moment des appels téléphoniques du cycle 2 les intervieweurs ne pouvaient pas demander le répondant par son nom. Ils utilisaient plutôt un descripteur, en l'occurrence l'âge et le sexe. Dans l'étude 2, on demandait à chaque répondant du cycle 1 un descripteur de nom, qu'on utilisait pour tenter d'entrer de nouveau en contact avec lui au moment du cycle 2. Dans l'étude 1, la perte d'effectifs (c'est-à-dire la proportion de répondants du cycle 1 non réinterviewés au cycle 2) a été de 47%, alors que dans l'étude 2 elle a été de 43%: l'écart entre les taux de perte d'effectifs a donc été relativement faible. Dans les deux enquêtes, la perte d'effectifs avait un rapport significatif avec l'âge, la race, le niveau de scolarité et le revenu. L'article examine l'alternative qui se pose en ces termes: réduire au minimum ou bien la perte d'effectifs ou bien la réaction de l'enquête, c'est-à-dire l'effet que peut avoir sur ses réponses le fait qu'il sera réinterviewé, deux facteurs qui peuvent influencer sur l'erreur totale de l'enquête. L'article contient aussi des suggestions sur les moyens de réduire la perte d'effectifs dans les enquêtes par panel par composition aléatoire.

MOTS CLÉS: Pertes d'effectifs; composition aléatoire; enquêtes téléphoniques.

## 1. INTRODUCTION ET REVUE DES TRAVAUX ANTÉRIEURS

Depuis quelques décennies, le problème de la perte d'effectifs dans les enquêtes par panel n'a reçu qu'une attention occasionnelle dans les travaux sur les méthodes d'enquête. Les articles publiés traitaient soit de moyens de réduire au minimum la perte d'effectifs dans les enquêtes par panel (voir par exemple Droegge et Crambert 1965; Crider, Willlets et Bealer 1971; McAllister, Goe et Bulter 1973; Freedman, Thornton et Camburn 1980; et Burgess 1989), soit de techniques statistiques pouvant servir à corriger les effets de la perte d'effectifs. (Voir par exemple Lehen et Koch 1974; Hausmann et Wise 1979; Winer 1983; et Lepkowski 1989.) Peu d'articles ont parlé de l'importance numérique et de la nature de la perte d'effectifs. Moins nombreux encore sont ceux qui ont offert sur les échantillons *aléatoires* des conclusions pouvant donner à d'autres chercheurs le moyen de savoir à quoi s'attendre dans des enquêtes portant sur l'ensemble de la population. L'article de Sobol (1959), qui étudiait la perte d'effectifs observée dans une enquête par panel à cinq cycles portant sur l'évolution des attitudes économiques, est une exception. Lors du cycle 1, qui a eu lieu en 1954, on a interviewé un échantillon probabiliste de la population urbaine hors établissement des États-Unis ( $n = 1,150$ ). Les cycles suivants ont eu lieu six, douze, dix-huit et trente-trois mois plus tard. Par rapport au premier échantillon, les pertes d'effectifs aux cycles ultérieurs ont été respectivement de 17, 26, 29 et 39%.

<sup>1</sup> Paul J. Lavrakas, Richard A. Settersten, Jr. et Richard A. Maier, Jr., Northwestern University Survey Laboratory 625 Haven St., Evanston IL 60208 - 4150 USA.





- RAO, P.S.R.S., KAPLAN, J., et COCHRAN, W.G. (1981). Estimators for the one-way random effects model with unequal error variances. *Journal of the American Statistical Association*, 76, 89-96.
- SCOTT, A.J., et SMITH, T.M.F. (1977). The application of time series methods to the analysis of repeated surveys. *Revue Internationale de Statistique*, 45, 13-28.
- SHIMIZU, I.M. (1987). Specifications for the redesigned NHDS sample. National Center for Health Statistics, rapport technique.
- VATES, F., et COCHRAN, W.G. (1938). The analysis of groups of experiments. *Journal of Agricultural Sciences*, 28, 556-580.

Pour estimer  $\sigma^2$  et  $\nu$ , nous nous sommes servis de la méthode des sommes de carrés non pondérées. On peut aussi examiner les effets de l'utilisation de l'analyse de variance et de l'EONBM à cette fin. Toutefois, l'étude de Rao *et al.* (1981) a montré que les différentes méthodes d'estimation de  $\sigma^2$  peuvent n'avoir pas d'effet sensible sur l'estimation de  $\mu$  ou sur l'erreur-type de cette estimation.

D'autres recherches sont nécessaires pour déterminer l'effet des différentes méthodes d'estimation des variances sur l'estimateur empirique de Bayes pour  $\mu$ .

Nous avons remplacé une estimation négative de  $\sigma^2$  par une quantité positive faible. Comme on peut le voir, cette correction peut produire une erreur-type faible pour les méthodes VC et EB et donner une idée trop optimiste des estimations de  $\mu$  et de  $\mu'$ . Il faudrait que ce problème soit étudié plus en profondeur.

Nous avons fait l'hypothèse d'un modèle linéaire pour la proportion. La transformation logit ou probit peut être utilisée avant d'appliquer ce modèle. Cependant, les estimations tirées de ces transformations ne sont justifiées que si la taille de la population et celle de l'échantillon sont toutes deux grandes. Les estimations proposées dans cet article peuvent être faites par des utilisateurs des secteurs public ou privé au moyen d'un logiciel d'usage courant.

Dans cet article, nous avons étudié un moyen d'améliorer les estimations relatives à chacun des hôpitaux. On peut aussi obtenir des estimations nationales pour un caractère d'échantillonnage comme l'infarctus aigu du myocarde ou les troubles mentaux en pondérant convenablement les estimations obtenues ci-dessus par l'inverse des probabilités de sélection des hôpitaux. Une telle méthode devrait améliorer la précision des estimations nationales. Comme l'ont suggéré par exemple Blight et Scott (1973) et Scott et Smith (1977), des méthodes de séries chronologiques comme ARMMI peuvent être employées pour estimer les proportions et les valeurs totales. Ces méthodes produiront des modèles différents selon les caractères étudiés. De plus, les logiciels que l'on peut actuellement se procurer et qui utilisent une telle approche supposent que la taille de la population est grande, que les variances de l'erreur sont égales et que la taille de l'échantillon est la même pour toutes les périodes. De telles hypothèses ne sont pas vérifiées dans le problème qui fait l'objet de cet article. Comme nous l'avons dit dans la section 1, les méthodes TR, VC et EB peuvent aussi être employées lorsqu'il y a non-réponse certaines années.

## REMERCIEMENTS

Les auteurs désirent remercier le rédacteur associé et le lecteur du manuscrit pour leurs commentaires utiles.

## BIBLIOGRAPHIE

BEAN, J. A. (1987). NHDS variance and covariance estimation of year to year differences. National Center for Health Statistics, rapport technique.

BLIGHT, B. J. N., et SCOTT, A. J. (1973). A stochastic model for repeated surveys. *Journal of the Royal Statistical Society, Series B*, 35, 61-68.

CARROLL, R. J., et RUPERT, D. (1988). *Transformation and Weighing in Regression*. New York: Chapman and Hall.

COCHRAN, W. G. (1954). The combination of estimates from different experiments. *Biometrics*, 10, 101-129.

RAO, C. R. (1972). Estimation of variance and covariance components in linear models. *Journal of the American Statistical Association*, 67, 112-115.



En utilisant les estimations  $\hat{\sigma}_a^2$ ,  $\hat{\sigma}_i^2$  et  $\hat{\mu}$ , (21) peut s'écrire:

$$B_i' = (1 - \hat{a}_i') y_i' + \hat{a}_i \mu, \tag{23}$$

où  $\hat{a}_i' = \hat{y}_i' / (\hat{\sigma}_a^2 + \hat{y}_i')$ . On peut appeler cet estimateur l'estimateur empirique de Bayes (EB). À noter que  $\hat{\mu}$  découle de (17) avec  $\hat{\sigma}_a^2$  et  $\hat{y}_i'$ . On peut obtenir la variance de cet estimateur au moyen de (22) en remplaçant  $\sigma_a^2$  et  $y_i'$  par leur estimation. Pour obtenir l'estimateur de Bayes et sa variance, nous avons estimé  $\sigma_a^2$  et  $y_i'$  en utilisant la méthode des sommes de carrés non pondérées, décrite à la section précédente.

### 7. EFFICACITÉ DES ESTIMATEURS

Pour chacun des vingt hôpitaux, nous avons effectué les estimations de  $P_i$  pour 1986 en utilisant chacune des méthodes décrites dans les sections qui précèdent. Comme nous l'avons mentionné précédemment, les valeurs de  $P_i$  pour l'ensemble de la population ne sont pas connues; nous avons donc calculé les erreurs-types correspondant aux différentes méthodes en substituant à  $P_i$  la proportion  $P_i$  observée dans l'échantillon. Comme les tailles d'échantillon  $n_i$  ne sont pas petites, on peut s'attendre à ce que les biais d'estimation de la variance et de l'erreur-type des différents estimateurs soient faibles.

Pour les trois hôpitaux, les estimations de  $P_i$  et les erreurs-types correspondant aux différentes méthodes sont données aux tableaux 2 et 3 pour l'infarctus aigu du myocarde et les troubles mentaux.

Comme ces tableaux l'indiquent, les erreurs-types des méthodes TR, VC et EB sont plus petites que l'erreur-type de la proportion d'échantillon. Comme pouvait le prévoir, l'utilisation des données des périodes précédentes a aidé à réduire l'erreur-type de l'estimation pour la dernière période.

Les méthodes VC et EB donnent toutes deux une erreur-type inférieure à celle de la méthode TR. Cependant, cette dernière méthode ne nécessite pas d'estimation de  $\sigma_a^2$ . Nous avons pu constater que l'erreur-type de l'estimateur de tendance était généralement inférieure de plus de la moitié à celle de la proportion d'échantillon.

La méthode EB produit une erreur-type inférieure à celle de la méthode VC, comme prévu. À noter que VC estime la proportion d'ensemble, tandis que EB estime la proportion de la distribution conditionnelle. L'erreur-type de EB prend une valeur proche de celle de la proportion d'échantillon si la taille de l'échantillon est grande.

Il est intéressant d'observer que dans les tableaux 2 et 3 la différence entre les estimations VC et EB est négligeable, pour l'infarctus aigu du myocarde aussi bien que pour les troubles mentaux. La raison en est que  $\hat{a}_i'$  est proche de l'unité, ce qui indique que  $\sigma_a^2$  est petit par rapport à  $y_i'$ .

Il est possible d'obtenir des estimations du nombre total de cas en 1986, avec les erreurs-types correspondantes, en multipliant les estimations des proportions des tableaux 2 et 3 par le nombre correspondant  $N_i$  de sorties d'hôpital, inscrit au tableau 1.

### 8. ANALYSE

Tels que décrits dans la section précédente, les résultats de cette recherche montrent qu'il est avantageux d'utiliser les méthodes TR, VC ou EB pour l'estimation des proportions et des valeurs totales correspondant au dernier passage d'une enquête.

Pour estimer les erreurs-types des différentes méthodes, nous avons utilisé les proportions d'échantillon. D'autres recherches sont nécessaires pour étudier le biais et l'erreur quadratique moyenne de ces erreurs-types.

Pour obtenir la moyenne (17) et sa variance (18), nous avons remplacé  $\sigma_i^2$  par son estimation  $\hat{P}_i(1 - \hat{P}_i)$ . Pour estimer  $\sigma_i^2$ , nous avons des méthodes telles que l'analyse de variance et l'estimation quadratique non biaisée de norme minimale (EQBNM). L'EQBNM dépend des valeurs *a priori*  $r_i$  de  $(\sigma_i^2/\sigma^2)$ . Nous décrivons ci-dessous une méthode apparentée, dite méthode des sommes de carrés non pondérées (SCNP), qui ne dépend pas de  $r_i$ . P.S.R.S. Rao, Kaplan et Cochran (1981) ont montré que cette méthode fournit des estimations de  $\sigma_i^2$  comparables à celles obtenues par l'analyse de variance ou l'EQBNM, sauf si  $n_i$  ou  $r_i$  est très petit. Les calculs qu'entraîne la méthode des sommes de carrés non pondérées sont moins lourds que pour l'estimation quadratique non biaisée de norme minimale. Avec  $y^* = (\sum y_i)/T$ , on tire de (15):

$$E[\sum (y_i - y^*)^2] = (T - 1)\sigma^2 + (T - 1)(\sum v_i)/T, \quad (19)$$

où  $v_i = (N_i - n_i)P_i(1 - P_i)/(N_i - 1)n_i$ . L'estimateur par sommes de carrés non pondérées de  $\sigma^2$  est:

$$\hat{\sigma}^2 = \sum (y_i - y^*)^2 / (T - 1) - (\sum v_i)/T, \quad (20)$$

où  $\hat{v}_i = (1 - \hat{f}_i)\hat{P}_i(1 - \hat{P}_i)/(n_i - 1)$ . Si  $N_i$  est grand par rapport à  $n_i$ , on peut remplacer la fraction de sondage  $f_i$  par zéro. Nous avons estimé  $U_i$  à l'aide de (16) en estimant  $\sigma_i^2$  à l'aide de (20) et en estimant le deuxième terme par  $\hat{v}_i$ . En utilisant cette estimation de  $U_i$ , nous avons estimé  $\mu$  à l'aide de (17) et sa variance à l'aide de (18). Si  $\sigma_i^2$  est beaucoup plus grand que  $v_i$ , l'estimateur  $\hat{\mu}$  de (17) sera proche de  $y^*$ . Dans ce cas, on peut s'attendre à ce que l'estimation de  $U_i$  décrite ci-dessus n'ait presque aucun effet sur  $\hat{\mu}$ . Comme  $\hat{\mu}$  dépend seulement de la valeur relative de  $U_i$ , on peut considérer que cette conclusion sera valide même quand  $\sigma_i^2$  n'est pas beaucoup plus grand que  $v_i$ . Ainsi, on peut considérer que l'estimation de  $U_i$  ne produira qu'un biais négligeable pour  $\hat{\mu}$ . Comme on sait, toutes les méthodes d'estimation sans biais de  $\sigma_i^2$  peuvent produire des estimations négatives. Dans ce cas, nous avons utilisé le procédé habituel, qui consiste à substituer une petite quantité positive à l'estimation négative. Dans Rao *et al.* (1981), il a été démontré que, sauf si  $\sigma_i^2$  est très petit, cette correction ne produit qu'un biais négligeable pour  $\hat{\sigma}^2$  et une augmentation peut sensiblement de son erreur-type. De plus, nous avons constaté que, sauf le cas où  $\sigma_i^2$  est petit, on a montré que la différence entre les erreurs quadratiques moyennes de  $\hat{\mu}$  obtenues en estimant  $U_i$  par la méthode des sommes de carrés non pondérées et par les autres méthodes d'estimation de  $U_i$  était négligeable.

## 6. ESTIMATEUR DE BAYES

La discussion du début de la section 5 donne à penser qu'il est possible de supposer que  $\mu_i$  a une distribution *a priori* de moyenne  $\mu$  et de variance  $\sigma_i^2$ . Dans l'hypothèse où, pour  $N_i$  grand, la distribution de  $y_{ij}$  est normale et qu'elle est de moyenne  $\mu_i$  et de variance  $\sigma_i^2$ , et que la distribution *a priori* est elle aussi normale, l'estimateur de Bayes de  $\mu_i$  est:

$$B_i = E(m_i | y_i) = (1 - a_i)y_i + a_i\mu, \quad (21)$$

où  $a_i = v_i/(\sigma_i^2 + v_i)$ . L'expression qui donne  $v_i$  est la même que celle présentée à la section précédente.

Pour un  $y_i$  donné, la variance de cet estimateur est donnée par:

$$V(B_i) = \frac{1}{1 + (1/\sigma_i^2) + (1/v_i)}. \quad (22)$$

La variance de l'estimateur ci-dessus est donnée par:

$$V(\hat{\mu}_i) = \frac{\sum W_i}{1} + \frac{\sum W_i(x_i - \bar{x})^2}{(x_i - \bar{x})^2}.$$

(13)

Nous avons estimé cette variance en remplaçant  $W_i$  par  $w_i$ . Le biais de l'estimateur ainsi obtenu sera petit pour  $n_i$  grand.

Dans cet article, pour les fins de l'illustration, nous prendrons  $i = (1, 2, \dots, 10)$ , et donc  $T = 10$ . Pour 1986, nous avons estimé la proportion d'un caractère d'échantillonnage et son erreur-type à l'aide des expressions (12) et (13) en prenant  $x_i = 10$ .

5. MODÈLE DES COMPOSANTES DE LA VARIANCE

Un examen des proportions observées pour l'infarctus aigu du myocarde et les troubles mentaux dans les vingt hôpitaux étudiés au cours des dix années considérées n'a pas permis de conclure à une tendance linéaire ou non linéaire bien définie. Pour tous ces hôpitaux, le type de variation des chiffres recueillis ressemblait passablement à celui des trois hôpitaux dont les valeurs sont présentées aux figures 1 et 2. Ces observations ont indiqué que la proportion de l'infarctus aigu du myocarde ou des troubles mentaux pour l'année considérée pouvait être obtenue en combinant l'information de l'ensemble des dix années. À cette fin, il est possible d'utiliser le modèle unidimensionnel avec effets aléatoires.

Le modèle formulé dans l'expression (8) peut s'écrire:

$$y_{ij} = \mu + (\mu_i - \mu) + \epsilon_{ij}$$

$$= \mu + \alpha_i + \epsilon_{ij}.$$

(14)

Si l'on considère que  $\mu_i$  est tiré au hasard d'une population de moyenne  $\mu$ , l'effet aléatoire  $\alpha_i$  sera de moyenne zéro et variance  $\sigma_a^2$ . On fait l'hypothèse qu'il est indépendant de  $\epsilon_{ij}$ . La moyenne de l'échantillon (proportion) peut maintenant s'écrire:

$$y_i = \mu + \alpha_i + \epsilon_i,$$

(15)

où  $\epsilon_i$  est de moyenne zéro et de variance  $(1 - f_i) \sigma_a^2/n_i$ . De (15) on déduit:

$$V(y_i) = \sigma_a^2 + (N_i - n_i)\sigma_i^2/(N_i - 1)n_i = \frac{1}{U_i}.$$

(16)

L'estimateur par les MCP de  $\mu$  est:

$$\hat{\mu} = \frac{\sum U_i y_i}{\sum U_i}.$$

(17)

C'est l'estimateur des composantes de la variance (VC) et sa variance est donnée par:

$$V(\hat{\mu}) = 1/\sum U_i.$$

(18)



où  $a_i$  est le nombre de cas observés pour ce caractère parmi les  $n_i$  sorties d'hôpital qui constituent l'échantillon. La variance de  $\hat{P}_i$  et son estimateur sans biais sont donnés par:

$$V(\hat{P}_i) = \frac{N_i - n_i}{N_i} \frac{P_i(1 - P_i)}{n_i} \tag{6}$$

et

$$v(\hat{P}_i) = (1 - f_i) \frac{P_i(1 - P_i)}{n_i - 1}, \tag{7}$$

où  $f_i = n_i/N_i$ . À noter que  $\hat{P}_i$  est la même chose que  $\hat{y}_i = \sum_{j=1}^n y_{ij}/n_i$ .

#### 4. TENDANCE LINÉAIRE

Les valeurs observées dans l'échantillon,  $y_{ij}$ ,  $j = (1, 2, \dots, n_{ij})$  peuvent s'écrire:

$$y_{ij} = \mu_i + \epsilon_{ij}, \tag{8}$$

où  $\mu_i$  est la moyenne du  $i^{\text{ème}}$  hôpital à la  $i^{\text{ème}}$  période, et  $\epsilon_{ij}$  est l'erreur aléatoire d'espérance 0 et de variance  $\sigma_i^2 = P_i(1 - P_i)$ . Comme les échantillons sont tirés de façon indépendante d'une année à l'autre, il n'y a pas de corrélation entre les erreurs  $\epsilon_{ij}$  d'une année à l'autre. En faisant l'hypothèse d'une tendance linéaire, la moyenne de l'échantillon peut s'écrire:

$$\bar{y}_i = \alpha + \beta x_i + \epsilon_i, \tag{9}$$

où  $x_i = i$  et  $\epsilon_i = \sum_{j=1}^{n_i} \epsilon_{ij}/n_i$ . De plus,  $V(\epsilon_i) = (N_i - n_i)\sigma_i^2/(N_i - 1)n_i = 1/W_i$ . On peut remarquer qu'avec la notation zéro-un  $\bar{y}_i$  est la même chose que  $\hat{P}_i$ . Les estimateurs par les MCP de  $\beta$  et  $\alpha$  sont:

$$\hat{\beta} = \frac{\sum W_i(x_i - \bar{x})(\bar{y}_i - \bar{y})}{\sum W_i(x_i - \bar{x})^2} \tag{10}$$

et

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}, \tag{11}$$

où  $\bar{x} = \sum W_i x_i / \sum W_i$  et  $\bar{y} = \sum W_i \bar{y}_i / \sum W_i$ .

L'estimateur de  $\mu_i$  est:

$$\hat{\mu}_i = \hat{\alpha} + \hat{\beta}x_i$$

$$= \hat{y} + \hat{\beta}(x_i - \bar{x}). \tag{12}$$

C'est l'estimateur de tendance (ET). Des estimateurs de ce type ont déjà été étudiés pour des populations infinies; voir par exemple Carrol et Rupert (1988).

Nous avons tiré l'estimateur de  $\mu_i$  de cette expression en remplaçant  $W_i$  par  $w_i = (1 - f_i)\hat{\sigma}_i^2/n_i$ , où  $\hat{\sigma}_i^2 = \hat{P}_i(1 - \hat{P}_i)$ . Si l'on peut faire l'hypothèse que, pour  $N_i$  grand, la distribution de  $y_{ij}$  est normale,  $\hat{y}_i$  sera indépendant de  $w_i$ . Dans ce cas, l'expression (12) demeure un estimateur sans biais de  $\mu_i$ . Même si l'hypothèse de normalité n'est pas vérifiée, on peut montrer que  $w_i$  tend vers  $W_i$  pour  $n_i$  grand et que, par conséquent, l'expression (12) qui contient les pondérations estimées tend vers  $\mu_i$  pour  $n_i$  grand.

Tableau 3  
Estimations des proportions pour 1986 et erreurs-types  
(chiffres du bas) pour les troubles mentaux (TM)

Hôpital	Proportion de l'échantillon	Estimation par tendance	Estimation par les composantes de la variance	Estimation bayésienne
1	.1326	.1431	.1292	.1292
	.0205	.0115	.0060	.0010
2	.0323	.0437	.0500	.0427
	.0087	.0056	.0039	.0057
3	.0504	.0496	.0534	.0523
	.0080	.0049	.0032	.0048

2. NOTATION

Dans cette section, nous présentons la notation pour une unité primaire d'échantillonnage sélectionnée. Soit  $y_{itj}$  la  $j^{ième}$  observation sur une sortie d'hôpital sélectionnée pour un caractère tel que le nombre de cas chirurgicaux au temps  $t = (1, 2, \dots, T)$  dans le  $i^{ième}$  hôpital,  $i = (1, 2, \dots, K)$ , qui a  $N_{it}$  sorties. Il est à noter que  $K$  peut varier d'une année à l'autre lorsque certains hôpitaux omettent de répondre ou que de nouveaux hôpitaux sont ajoutés à l'échantillon.

Le total et la moyenne au temps  $t$  sont donnés par:

(1) 
$$Y_{it} = \sum_{N_{it}}^1 y_{itj}$$

et

(2) 
$$Y_{it} = Y_{it}/N_{it}.$$

Le total et la moyenne de l'échantillon de taille  $n_{it}$  tiré des  $N_{it}$  sorties d'hôpital sont donnés par:

(3) 
$$y_{it} = \sum_{n_{it}}^1 y_{itj}$$

(4) 
$$y_{it} = y_{it}/n_{it}.$$

Pour estimer le total et la proportion d'un caractère précis, soit  $y_{itj} = 1$ , si l'observation a trait à ce caractère, et zéro autrement. Avec cette notation, le total et la proportion d'un caractère au temps  $t$  peuvent s'écrire  $A_{it}$  et  $P_{it} = A_{it}/N_{it}$ . À noter que  $P_{it}$  est la même chose que  $Y_{it}$ . Dans les quatre sections suivantes, pour des raisons de commodité, nous supprimons l'indice  $i$  et décrivons les estimateurs pour un hôpital donné.

3. PROPORTION DE L'ÉCHANTILLON

Un estimateur sans biais de la proportion  $P_{it}$  pour un caractère d'échantillonnage comme l'infarctus aigu du myocarde ou les troubles mentaux est donné par:

(5) 
$$P_{it} = a_{it}/n_{it},$$

La deuxième méthode utilise le modèle unidimensionnel avec effets aléatoires, à variances inégales, pour combiner l'information des différentes années. Yates et Cochran (1938) et Cochran (1954) ont proposé ce procédé pour combiner l'information provenant d'expériences effectuées à des endroits ou à des moments différents. Bien que l'analyse de variance soit depuis assez longtemps utilisée à cette fin, C.R. Rao (1970) a proposé l'estimation quadratique non biaisée de norme minimale (EQNBM) et démontré ses avantages. P.S.R.S. Rao, Kaplan et Cochran (1981) ont examiné la valeur relative de l'analyse de variance, de l'EQNBM et de plusieurs autres méthodes apparentées. Nous avons employé les procédés d'estimation correspondant à ces méthodes. L'estimation de la proportion obtenue par l'un ou l'autre de ces procédés est une combinaison pondérée des estimations pour les différentes périodes. Les poids dépendent de la variance entre les périodes aussi bien que de la variance à l'intérieur des périodes. La troisième méthode utilise la méthode empirique de Bayes pour estimer les proportions de la période en cours.

Nous notons les trois méthodes indiquées ci-dessus TR, VC et EB. La section 2 présente la notation. La section 3 donne l'estimateur de la proportion fondé sur l'échantillon avec la variance de cet estimateur. Les sections 4, 5 et 6 présentent chacune des trois méthodes mentionnées ci-dessus et donnent pour chacune l'expression de son erreur-type. Nous avons utilisé ces expressions pour établir, pour l'année 1986, les proportions d'échantillon ainsi que les trois types d'estimation et leur erreur-type pour vingt des hôpitaux étudiés dans la NHDS. Pour les trois hôpitaux mentionnés plus haut, ces trois estimations sont présentées au tableau 2 pour l'infarctus aigu du myocarde et au tableau 3 pour les troubles mentaux. La section 7 présente les résultats pour l'ensemble de l'étude de la NHDS. La dernière section examine les résultats et présente des sujets de recherche.

Pour le problème étudié dans cet article, le tirage des échantillons se fait de façon indépendante pour les différentes périodes. De plus, les proportions qui existaient dans la population pendant les périodes précédentes ne sont pas connues. C'est pourquoi on ne peut pas utiliser les méthodes habituelles – la méthode des quotients et la régression – pour améliorer la précision des estimateurs pour la période considérée. Pour les mêmes raisons, on ne peut pas utiliser ici les méthodes d'estimation que proposent diverses publications pour les sondages avec renouvellement. Malgré ces difficultés, il est possible de se servir des trois méthodes qui font l'objet de cet article pour estimer avec un degré élevé de précision les paramètres de la population. Lorsque des chiffres globaux pour les différentes périodes sont disponibles, les utilisateurs du secteur public comme du secteur privé peuvent établir ces estimations, avec leur erreur-type, sans trop de difficulté. On peut même utiliser ces méthodes quand il y a eu non-réponse certaines années, car il y a des hôpitaux qui ne communiquent pas de renseignements certaines années.

Tableau 2

Estimations des proportions pour 1986 et erreurs-types (chiffres du bas)  
pour l'infarctus aigu du myocarde (IAM)

Hôpital	Proportion de l'échantillon	Estimation par tendance	Estimation par les composantes de la variance	Estimation bayésienne
1	.0152	.0196	.0224	.0224
	.0070	.0046	.0026	.0003
2	.0108	.0162	.0204	.0203
	.0048	.0036	.0031	.0003
3	.0321	.0319	.0304	.0309
	.0060	.0038	.0028	.0037



Figure 1. Proportions pour l'IAM, 1977 à 1986

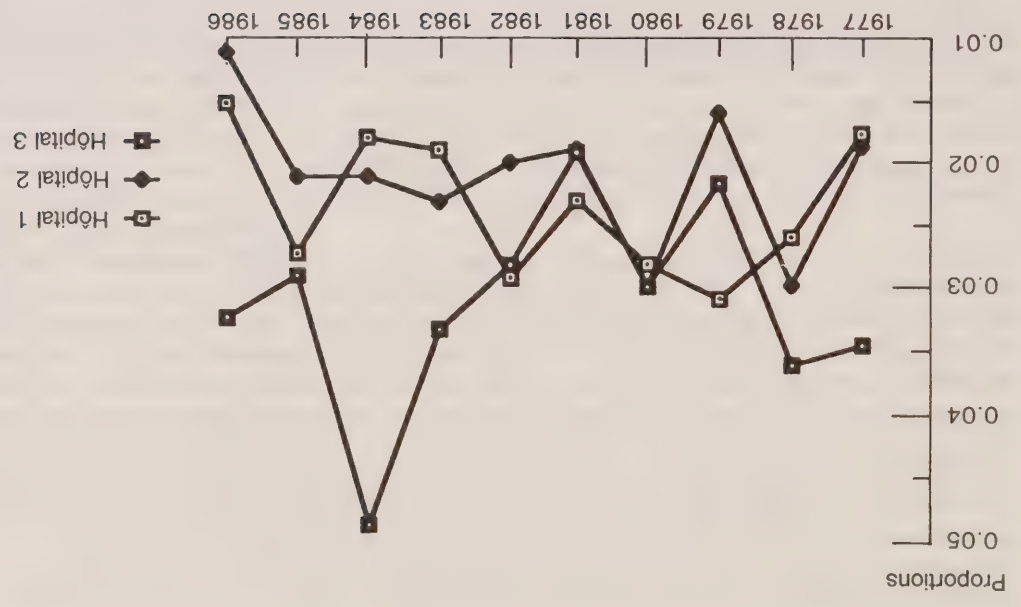


Figure 2. Proportions pour les TM, 1977 à 1986

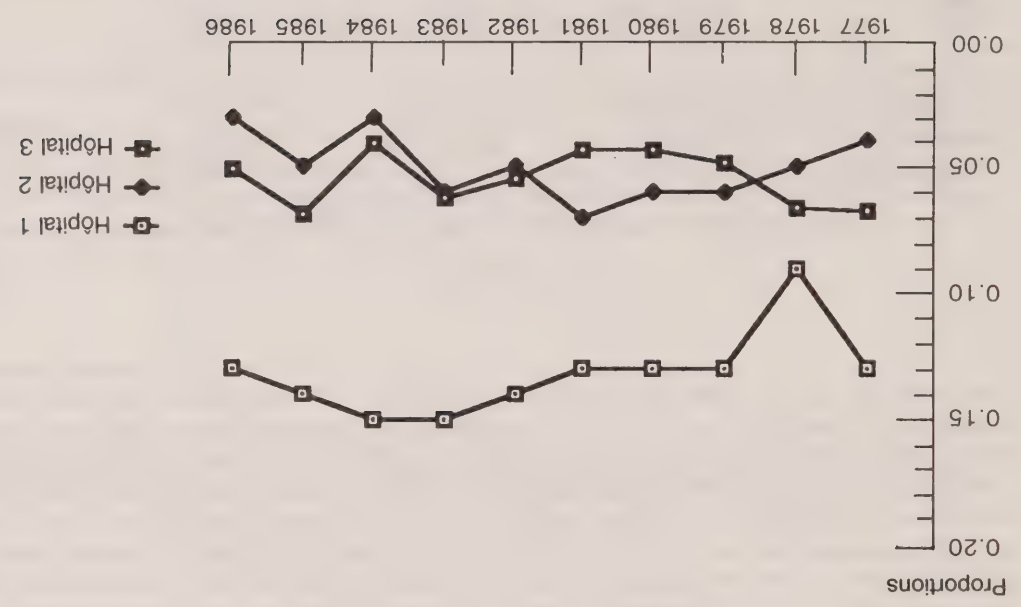


Tableau 1  
Données de la National Health Discharge Survey, 1977 à 1986  
Valeurs totales et proportions observées dans  
des échantillons pour l'infarctus aigu du myocarde (IAM)  
et les troubles mentaux (TM) dans trois hôpitaux

Année	Nombre de sorties d'hôpital N	Nombre de sorties d'hôpital n	IAM		MDS	
			Total	Proportion de l'échantillon	Total	Proportion de l'échantillon

1977	9,416	276	5	.018	37	.13
1978	10,234	266	7	.026	24	.09
1979	9,354	294	9	.031	39	.13
1980	10,372	327	9	.028	41	.13
1981	10,712	342	8	.023	45	.13
1982	10,683	309	9	.029	43	.14
1983	10,935	360	7	.019	46	.15
1984	10,090	330	6	.018	50	.15
1985	10,431	297	8	.027	41	.14
1986	10,247	264	4	.015	35	.13

1977	6,720	474	9	.019	18	.04
1978	6,710	470	14	.030	25	.05
1979	6,970	495	8	.016	28	.06
1980	6,794	466	14	.030	29	.06
1981	7,055	486	9	.019	34	.07
1982	6,265	442	9	.020	24	.05
1983	6,234	442	10	.023	28	.06
1984	6,221	439	9	.021	15	.03
1985	6,063	375	8	.021	19	.05
1986	5,781	371	4	.011	12	.03

1977	6,400	606	21	.0347	41	.0677
1978	6,286	635	23	.0362	42	.0661
1979	6,494	554	12	.0217	27	.0487
1980	6,813	571	17	.0298	25	.0438
1981	7,430	729	14	.0192	32	.0439
1982	7,267	712	20	.0281	39	.0548
1983	7,110	694	23	.0331	43	.0620
1984	7,268	718	35	.0487	29	.0404
1985	6,716	657	19	.0289	45	.0685
1986	6,464	655	21	.0321	33	.0504

Les proportions d'échantillon du tableau 1 et des figures 1 et 2 ne doivent toutefois pas être utilisées pour formuler des conclusions sur l'augmentation ou la diminution de la fréquence de l'infarctus du myocarde ou des troubles mentaux dans l'ensemble de la population.

Dans cet article, nous examinons trois méthodes pour avant améliorer les estimations relatives à un hôpital donné en utilisant l'information des années précédentes aussi bien que celle de l'année considérée. Avec la première méthode, on obtient des estimations des proportions à l'aide de la tendance linéaire sur l'ensemble des années et de la méthode des moindres carrés pondérés. S'il y a une tendance significative – positive ou négative – pour ces années, cette méthode aura plus de précision que l'estimation tirée de l'échantillon de la dernière période. Si la tendance n'est pas prononcée, le gain de précision sera naturellement négligeable.

## Combinaison d'estimations provenant d'enquêtes

PODURI S.R.S. RAO et I.M. SHIMIZU<sup>1</sup>

### RÉSUMÉ

Pour estimer la proportion et la valeur totale d'un caractère d'échantillonnage pour le dernier passage d'une enquête, on utilise trois méthodes qui permettent de combiner les estimations indépendantes de ce caractère relatives à ce dernier passage et aux précédents. Dans la première méthode, on étudie la tendance sur l'ensemble des passages. Dans la deuxième, on emploie le modèle unidimensionnel avec effets aléatoires. Dans la troisième, on se sert de l'approche empirique de Bayes. Chacune de ces trois méthodes se révèle plus efficace que l'estimation par échantillon fondée sur les données du dernier passage seul. On discute des avantages et des limites de ces méthodes, toutes trois étant illustrées par des données de la National Health Discharge Survey (enquête nationale sur les sorties d'hôpital).

**MOTS CLÉS:** Tendance; moindres carrés pondérés; effets aléatoires; amélioration de l'estimation; biais; écarts quadratiques moyens.

### 1. INTRODUCTION

Plusieurs enquêtes nationales prélèvent des échantillons indépendants à des périodes successives. Dans cet article, l'information tirée des enquêtes précédentes est utilisée pour améliorer les estimations relatives à la dernière période. Pour illustrer notre propos, nous avons choisi la National Health Discharge Survey (NHDS), effectuée aux États-Unis. Dans cette enquête, dont la méthodologie a récemment été révisée, on utilise un plan d'échantillonnage à trois degrés, où les régions servent d'unités primaires d'échantillonnage au premier degré. Ce sont les hôpitaux et les sorties d'hôpital qui sont sélectionnées aux deuxième et troisième degrés respectivement. L'enquête recueille des renseignements sur différentes caractéristiques des patients comme l'âge, le sexe, les caractéristiques raciales, la durée du séjour, le diagnostic et les opérations chirurgicales et autres traitements. Les unités primaires d'échantillonnage et les hôpitaux sélectionnés pour cette enquête demeurent les mêmes pendant un certain nombre d'années. Des échantillons indépendants de sorties sont recueillis chaque année dans ces hôpitaux. Shimizu (1987) donne des renseignements supplémentaires sur la révision des méthodes de la NHDS.

Actuellement, pour un hôpital donné, les estimations des proportions des différents caractères pour l'année en cours se fondent uniquement sur les données de cette même année. Des estimations nationales sont obtenues en pondérant convenablement ces proportions par les valeurs inverses des probabilités de sélection des hôpitaux et des unités primaires d'échantillonnage. Toutefois, Bean (1987) a constaté que, pour la plupart des caractères, il y avait une certaine corrélation entre les estimations des différentes années. Par exemple, les proportions d'échantillon observées de 1977 à 1986 dans le cadre de la NHDS pour l'infarctus aigu du myocarde (IAM) et pour les troubles mentaux (TM) sont présentées pour trois hôpitaux au tableau 1 et illustrées aux figures 1 et 2. Un examen des proportions observées dans ces trois hôpitaux et dans dix-sept autres a montré que la prise en compte de données de périodes antérieures pouvait augmenter la précision des estimations pour l'année considérée.

<sup>1</sup> Poduri S.R.S. Rao, Department of Statistics, Hyilan 703, University of Rochester, Rochester NY, 14618 U.S.A., et I.M. Shimizu, National Center for Health Statistics, Office of Research and Methodology 1-68, 3700 East-West Highway, Hyattsville MD, 20782, U.S.A.



## BIBLIOGRAPHIE

- ALHO, J. (1990). Logistic regression in capture-recapture models. *Biometrics*, 46, 623-635.
- BURNHAM, P.K., et OVERTON, W.S. (1979). Robust estimation of population size when capture probabilities vary among animals. *Ecology*, 60, 927-936.
- FELLER, W. (1968). *An Introduction to Probability Theory and Its Applications*, (Vol. I, 3<sup>ème</sup> éd.). New York: Wiley.
- HUGGINS, R.M. (1989). On the statistical analysis of capture experiments. *Biometrika*, 76, 133-140.
- SEBER, G.A.F. (1982). *The Estimation of Animal Abundance*, (2<sup>ème</sup> éd.). New York: Griffin.
- SEKAR, C.C., et DEMING, W.E. (1949). On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association*, 44, 101-115.
- VAAKANEN, V., VASAMA, M., et ALHO, J. (1985). Occupational diseases in Finland in 1984. Reviews 11, Institute of Occupational Health, Helsinki.

faible ou si  $N$  est faible. En conclusion, il semble que la méthode de système dual donnera des résultats satisfaisants uniquement si les systèmes d'enregistrement sont plus ou moins non corrélés ou si l'hétérogénéité est observable. Dans ce dernier cas, nous pouvons recourir à la stratification, comme l'ont déjà proposé Sekar et Deming (1949), ou aux modèles de régression logistique, comme le proposent Huggins (1989) et Alho (1990), pour atténuer l'effet du biais de l'estimateur classique de la taille d'une population.

REMERCIEMENTS

L'auteur tient à remercier Bruce Spencer et un arbitre anonyme, qui ont permis par leurs commentaires d'améliorer la présentation de cet article. Une partie des résultats empiriques ont déjà été présentés au I<sup>lème</sup> congrès des pays nordiques sur la statistique mathématique, tenu à Uppsala, Suède, en juin 1986.

ANNEXE

**Démonstration du lemme 1.** Appliquons un développement de Taylor à  $\hat{N} = n_1 n_2 / m$  faisant

$$\hat{N} \approx \frac{N \bar{p}_1 \bar{p}_2}{\bar{p}_2} + \frac{p_{12}}{\bar{p}_2} (n_1 - N \bar{p}_1) + \frac{p_{12}}{\bar{p}_1} (n_2 - N \bar{p}_2) - \frac{p_{12}}{\bar{p}_1 \bar{p}_2} (m - N \bar{p}_{12}).$$

Nous avons alors

$$E \left[ \left( \hat{N} - \frac{N \bar{p}_1 \bar{p}_2}{\bar{p}_2} \right)^2 \right] \approx \left( \frac{\bar{p}_2}{\bar{p}_2} \right)^2 \text{Var}(n_1) + \left( \frac{\bar{p}_1}{\bar{p}_1} \right)^2 \text{Var}(n_2) + \left( \frac{\bar{p}_1 \bar{p}_2}{\bar{p}_2} \right)^2 \text{Var}(m) - \frac{\bar{p}_1 \bar{p}_2}{\bar{p}_2} \text{Cov}(n_1, m) - 2 \frac{\bar{p}_1 \bar{p}_2}{\bar{p}_2} \text{Cov}(n_2, m).$$

Suivant les hypothèses d'indépendance,  $\text{Var}(n_j) = N \bar{p}_j - S_j, j = 1, 2; \text{Var}(m) = N \bar{p}_{12} - S_3$ ,  $\text{Cov}(n_1, m) = -S_4 + N \bar{p}_{12}$ , et  $\text{Cov}(n_2, m) = -S_5 + N \bar{p}_{12}$ . En faisant les substitutions nécessaires dans la formule de l'erreur quadratique moyenne, on obtient le résultat recherché. **Démonstration de la proposition 1.** Faisons abstraction du terme négatif qui comprend  $S_3$  dans la formule du lemme 1. Comme  $0 < p_{ji} < 1$ , nous avons  $S_4 < N \bar{p}_{12}$ , et  $S_4 < S_1$ . Par conséquent,

$$2 \frac{\bar{p}_1 \bar{p}_2}{\bar{p}_2} S_4 < \frac{\bar{p}_3}{\bar{p}_1 \bar{p}_2} N \bar{p}_{12} + \frac{\bar{p}_2}{\bar{p}_2} S_1 + \left( \frac{\bar{p}_2}{\bar{p}_1 - \bar{p}_{12}} \right) \frac{\bar{p}_2}{\bar{p}_2} N \bar{p}_{12}.$$

De même,

$$2 \frac{\bar{p}_2 \bar{p}_2}{\bar{p}_2} S_5 < \frac{\bar{p}_3}{\bar{p}_1 \bar{p}_2} N \bar{p}_{12} + \frac{\bar{p}_2}{\bar{p}_2} S_2 + \left( \frac{\bar{p}_2}{\bar{p}_2 - \bar{p}_{12}} \right) \frac{\bar{p}_2}{\bar{p}_2} N \bar{p}_{12}.$$

En substituant ces bornes aux termes de l'expression du lemme 1, nous obtenons

$$\text{Var}(\hat{N}) > \frac{\bar{p}_2 \bar{p}_2}{\bar{p}_2} N + \frac{(\bar{p}_1 - \bar{p}_{12}) \bar{p}_2}{(\bar{p}_1 - \bar{p}_{12}) \bar{p}_2} N + \frac{\bar{p}_2}{(\bar{p}_2 - \bar{p}_{12}) \bar{p}_2} N.$$

En estimant  $N \bar{p}_j$  au moyen de  $n_j, j = 1, 2$ ; et  $N \bar{p}_{12}$  au moyen de  $m$  nous obtenons le résultat recherché.

cas déclarés annuellement à l'institut a varié de 4,000 à plus de 7,000 (soit de 0.2 à 0.4% de la population active occupée). Les principales catégories de maladies enregistrées par l'institut sont la perte d'audition due au bruit, les maladies causées par un travail répétitif ou monotone (épicondylite, bursite, ténosynovite) et les maladies de la peau (voir Vaaranen et coll. 1985). Le registre en question peut être vu comme un système d'enregistrement double car selon les règlements en vigueur, chaque cas de maladie doit être déclaré à l'institut par la compagnie d'assurance d'une part et par le médecin traitant d'autre part.

Il se peut que la probabilité de déclaration d'un cas dépende du diagnostic par exemple. En effet, d'après des données de 1981, nous avons ce qui suit: nombre de cas déclarés par les compagnies d'assurance,  $n_1 = 3,769$ ; nombre de cas déclarés par les médecins,  $n_2 = 3,053$  et nombre de cas déclarés par l'une et l'autre source,  $m = 1,591$ . Ainsi, l'estimation de système dual habituelle est  $N = 7,232$  et  $V_1^1 = 97$ ,  $V_2^2 = 222$ , et  $V_3^3 = 108.0$ . La ressemblance entre  $V_3$  et  $V_1$  est frappante. Si on stratifie les données selon quatre catégories de maladies (les trois catégories mentionnées plus haut et la catégorie "autres"), on obtient les estimations suivantes: - perte d'audition due au bruit:  $N = 2,230$ ,  $V_1^1 = 33.4$ ,  $V_2^2 = 47.2$ , et  $V_3^3 = 42.6$ ; - maladies causées par un travail répétitif ou monotone:  $N = 3,572$ ,  $V_1^1 = 201.4$ ,  $V_2^2 = 303.8$ , et  $V_3^3 = 204.2$ ; - maladies de la peau:  $N = 1,441$ ,  $V_1^1 = 30.9$ ,  $V_2^2 = 86.2$ , et  $V_3^3 = 37.5$ ; - autres maladies:  $N = 1,015$ ,  $V_1^1 = 32.7$ ,  $V_2^2 = 79.1$ , et  $V_3^3 = 37.2$ . En regroupant ces résultats, on obtient les estimations globales suivantes:  $N = 8,258$ ,  $V_1^1 = 209.0$ ,  $V_2^2 = 340.3$ , et  $V_3^3 = 215.2$ . Nous remarquons que les maladies causées par un travail répétitif ou monotone sont sous-déclarées dans une assez large mesure.

Nous avons poussé plus loin l'analyse en stratifiant les données selon le genre de maladie (4 classes), l'assureur (11 classes) et le groupe d'activité économique (7 classes). On pourrait croire *a priori* que ces facteurs influent sur les probabilités de déclaration. Or, la stratification n'a pas modifié de façon notable l'estimation ponctuelle. Par contre, elle a fait s'accroître de plus du tiers les écarts types estimés à cause vraisemblablement de l'existence de très petites strates. Cette analyse nous permet de conclure qu'en ce qui concerne cette application, le biais lié au diagnostic représente la principale source d'erreur dans l'estimateur classique.

Nous avons aussi analysé les données à l'aide d'une méthode de régression logistique, laquelle permet de tenir compte de l'hétérogénéité observable d'une population qui résulte de variables explicatives discrètes et continues. Ainsi, nous avons pu observer que pour chaque catégorie de maladies, l'âge avait une incidence sur la probabilité de déclaration pour une source mais non pour l'autre. Les estimations ponctuelles demeuraient donc les mêmes et la conclusion concernant le rôle du diagnostic ne pouvait être réfutée (Alho 1990).

## 5. ANALYSE

Les résultats théoriques montrent que l'estimateur de la variance habituel,  $V_1$  est conservatif lorsque les deux systèmes d'enregistrement sont corrélés négativement ou qu'ils sont indépendants l'un de l'autre. Par le principe de continuité,  $V_1$  peut aussi être conservatif lorsqu'il y a une corrélation positive mais faible. Lorsqu'il existe une forte corrélation positive,  $V_1$  produit des sous-estimations. Nous avons donc proposé un autre estimateur,  $V_2$ , qui est conservatif dans des conditions d'hétérogénéité quelconque, or, cet estimateur semble trop conservatif lorsqu'on le compare à  $V_3$  qui, lui, est assurément conservatif dans des conditions d'hétérogénéité gaussienne. La ressemblance entre  $V_3$  et  $V_1$  donne à penser qu'en pratique,  $V_1$  peut être assez robuste à l'égard de l'hétérogénéité d'une population.

Même en utilisant l'estimateur conservatif  $V_2$  dans notre exemple empirique, nous n'aurions pas réussi à compenser le biais de l'estimateur classique. Cela était peut-être prévisible car le biais de  $N$  et le degré de surestimation produite par  $V_2$  sont tous deux d'ordre  $N$ . L'utilisation de  $V_2$  n'accroîtrait donc la largeur d'un intervalle de confiance que par un facteur d'ordre  $N^{1/2}$ . Par conséquent,  $V_2$  peut compenser le biais de  $N$  uniquement si ce biais est



Quelle valeur devons-nous accorder à l'hypothèse des moments gaussiens? De toute évidence, les probabilités de saisie ne peuvent suivre une distribution strictement gaussienne car avec ce genre de distribution, il y a toujours une partie de la masse de probabilité qui se trouve à l'extérieur de l'intervalle unité. Par ailleurs, supposons que nous obtenions les  $p_{ji}$  au moyen de la formule  $\logit(p_{ji}) = a_j + b_j X_{ji}$ , où les paires  $(X_{1i}, X_{2i})$  forment un échantillon provenant d'une distribution normale bidimensionnelle de moyenne nulle, de variance unité et de corrélation  $\rho$ . Si nous avons les relations  $a_j = \logit(\mu_j)$ ,  $j = 1, 2$ , et  $b_j = \nu_j(1 + \mu_j^2)$ , l'hypothèse des moments gaussiens est approximativement juste. De fait, même la distribution des paires  $(p_{1i}, p_{2i})$  est, à peu de choses près, une distribution gaussienne bidimensionnelle. Examinons maintenant le bien-fondé de l'estimateur  $V_1$  selon l'hypothèse des moments gaussiens. Comme les probabilités doivent prendre une valeur entre 0 et 1,  $\mu_j$  sera aussi limitée au même intervalle. En outre, pour nous assurer que la plus grande partie de la masse de probabilité se trouve dans le carré-unité, supposons que  $0 < \nu_j \leq 1/2$ ,  $j = 1, 2$ . Si les valeurs de  $\mu_j$  sont proches de un, on peut choisir une borne supérieure beaucoup plus petite. Supposons maintenant que  $\rho \leq 0$ . On peut alors montrer que

$$-R \geq (\rho^2 \nu_1^4 \nu_2^2 + \rho^2 \nu_2^4 \nu_1^2) / (1 + \rho \nu_1 \nu_2)^4 > 0,$$

de sorte que  $V_1$  produit aussi une surestimation de  $\text{Var}(N)$  lorsque  $\rho \leq 0$ . Notons que, par le principe de continuité,  $V_1$  doit sûrement produire une surestimation de  $\text{Var}(N)$  pour des valeurs positives de  $\rho$ .

On peut montrer que  $R = R(\rho)$  est une fonction croissante de  $\rho$  à tout le moins pour les valeurs positives de  $\rho > 0$ . À la limite, nous avons

$$-R(\rho) \rightarrow (-2\nu_1\nu_2 - 2\nu_1^2\nu_2^2 + \nu_1^4\nu_2^2 + \nu_2^4\nu_1^2) / (1 + \nu_1\nu_2)^4,$$

lorsque  $\rho \rightarrow 1$ . Lorsque  $0 < \nu_j \leq 1/2$ ,  $j = 1, 2$ , la limite ci-dessus atteint sa valeur minimum à  $\nu_1 = \nu_2 = 1/2$ . Cette valeur minimum est  $-152/625 > -1/4$ . Par conséquent, pour  $\rho > 0$ ,  $V_1$  peut produire soit une sous-estimation soit une surestimation de  $\text{Var}(N)$ .

Concrètement, les résultats ci-dessus nous amènent aux conclusions suivantes. Premièrement, si  $\rho \leq 0$ , soit que  $N$  est convergent ou qu'il donne une surestimation de  $N_j$ ; de plus,  $V_1$  produit une surestimation de la variance, de sorte que nous pouvons calculer la limite supérieure d'un intervalle de confiance conservatif pour  $N$ . Lorsque  $\rho > 0$ ,  $N$  donne une sous-estimation de  $N$ . Si, en outre,  $\rho$  est petit,  $V_1$  produit une surestimation et il est alors possible de calculer la limite inférieure d'un intervalle de confiance conservatif pour  $N$ . Évidemment, il s'agit là de conditions plutôt particulières qui n'ont pas vraiment d'application pratique.

Selon le présent modèle, le biais asymptotique de  $V_1$  est plus grand que  $-N/4$  pour toutes les valeurs de  $\rho$ . Compte tenu de ce que le biais relatif asymptotique de  $\hat{N}$  est  $-\rho \nu_1 \nu_2 / (1 + \rho \nu_1 \nu_2) \geq -1/5$  dans le cas de la distribution gaussienne, nous pouvons déterminer un estimateur conservatif de la variance. Ainsi,  $5N/4 \geq N$  asymptotiquement. On a par exemple, comme estimateur conservatif de  $\text{Var}(N)$ ,  $V_3 = V_1 + 5N/16$ . Cet estimateur peut être beaucoup plus petit que  $V_2$ , ce qui donne une très grande valeur à l'hypothèse des moments gaussiens.

#### 4. EXEMPLE D'APPLICATION: DONNÉES SUR L'ENREGISTREMENT DES CAS DE MALADIE PROFESSIONNELLE

Pour avoir une idée de l'ordre de grandeur des biais dans la réalité, nous allons analyser des données sur l'enregistrement des cas de maladie professionnelle en Finlande. Ce pays tient un registre des maladies professionnelles depuis 1964. La tenue de ce registre est confiée à l'Institut de médecine du travail (traduction) à Helsinki. Depuis 1975, le nombre de nouveaux

3. HÉTÉROGÉNÉITÉ GAUSSIENNE

Nous allons maintenant analyser le cas particulier où les moments empiriques des paires  $(D_{1i}, D_{2i}), i = 1, \dots, N$ , concordent avec ceux d'une distribution normale (ou gaussienne) à deux variables. Cette analyse va nous permettre de définir avec beaucoup plus de précision que nous l'avons fait précédemment ce qu'est un estimateur conservatif de la variance. Supposons que

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N \left( \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} v_1^2 \mu_2^2 & \rho v_1 v_2 \mu_1 \mu_2 \\ \rho v_1 v_2 \mu_1 \mu_2 & v_2^2 \mu_1^2 \end{bmatrix} \right),$$

où  $|\rho| < 1$ , et  $0 < \mu_j < 1, j = 1, 2$ . Notons que  $v_j$  peut être considéré comme le coefficient de variation de la distribution de  $p_{ji}$ . Écrivons, comme avant,  $S_j = S_j/N$ , pour  $j = 1, \dots, 5$ . Intégrons ensuite les moments de la distribution normale bidimensionnelle à la formule du lemme 1 selon les équivalences suivantes:

$$\bar{p}_j = E[X_j] = \mu_j, j = 1, 2;$$

$$S_j = E[X_j^2] = \mu_j^2(1 + v_j^2), j = 1, 2;$$

$$p_{12} = E[X_1 X_2] = \mu_1 \mu_2(1 + \rho v_1 v_2);$$

$$\bar{S}_3 = E[X_1^2 X_2^2] = \mu_1^2 \mu_2^2(1 + v_1^2 + v_2^2 + 4\rho v_1 v_2 + (2\rho^2 + 1)v_1^2 v_2^2);$$

$$S_4 = E[X_1^2 X_2] = \mu_1^2 \mu_2(1 + 2\rho v_1 v_2 + v_1^2);$$

$$\bar{S}_5 = E[X_1 X_2^2] = \mu_1 \mu_2^2(1 + 2\rho v_1 v_2 + v_2^2).$$

Alors, par un calcul simple mais quelque peu fastidieux, nous pouvons démontrer la proposition suivante (détails exclus).

**Proposition 2.** Étant donné les hypothèses ci-dessus

$$\text{Var}(N) = A_1 A_2 + RN,$$

où

$$A_1 = N/(1 + \rho v_1 v_2)^2;$$

$$A_2 = [1 - (\mu_1 + \mu_2)(1 + \rho v_1 v_2) + \mu_1 \mu_2(1 + \rho v_1 v_2)^2]/[\mu_1 \mu_2(1 + \rho v_1 v_2)^2];$$

$$R = (2\rho v_1 v_2 + 3\rho^2 v_1^2 v_2^2 - \rho^2 v_1^4 v_2^2 - \rho^2 v_2^4 v_1^2 - v_1^2 v_2^2)/(1 + \rho v_1 v_2)^4.$$

À l'aide des résultats ci-dessus, nous pouvons calculer l'estimateur classique de la variance, estimateur qui produit une surestimation de la variance asymptotique, comme cela a été mentionné dans l'exemple 2.

$V_1 = n_1 n_2 u_1 u_2 / m^3$ . Notons tout d'abord que  $\{n_1 n_2 / m\} / A_1 \rightarrow 1$ , lorsque  $N \rightarrow \infty$ . De même,  $\{u_1 u_2 / m^2\} / A_2 \rightarrow 1$ . Cela démontre le corollaire de la proposition 2, à savoir  $(V_1 - \text{Var}(N)) / N \rightarrow -R$ , lorsque  $N \rightarrow \infty$ . Par exemple, si  $\rho = 0$ , alors  $-R = v_1^2 v_2^2$ , de sorte que  $V_1$  est un

où  $S_j = S_j/N$  pour  $j = 1, \dots, 5$ , et

$$S_1 = \sum_N p_2^{1i}, \quad S_2 = \sum_N p_2^{2i}, \quad S_3 = \sum_N p_2^{1i} p_2^{2i},$$
$$S_4 = \sum_N p_2^{1i} p_2^{2i}, \quad S_5 = \sum_N p_2^{1i} p_2^{2i}.$$

La démonstration de ce lemme est présentée en annexe. Nous constatons que, contrairement au biais de  $N_i$ , qui dépend uniquement des deux premiers moments des paires  $(p_{1i}, p_{2i})$ ,  $\text{Var}(N_i)$  dépend des quatre premiers moments. Dans des cas particuliers comme ceux décrits dans l'exemple 2 et la proposition 2, on peut simplifier la représentation.

**Exemple 1.** Supposons que les probabilités de saisie ne sont pas hétérogènes, c.-à-d.  $p_{ji} = p_j$ ,  $j = 1, 2$ . Alors,  $p_j = p_j$ ,  $j = 1, 2$ ;  $p_{12} = p_1 p_2$ ;  $S_j = p_j^2$ ,  $j = 1, 2$ ;  $S_3 = p_1^2 p_2^2$ ,  $S_4 = p_1^2 p_2^2$  et  $S_5 = p_1 p_2^2$ . Par conséquent, la variance asymptotique est  $\text{Var}(N_i) = N(1 - p_1 - p_2 + p_1 p_2)/(p_1 p_2) = Nq_1 q_2/(p_1 p_2)$ . Des estimateurs convergents de  $Np_1 p_2$  et de  $Np_j$  sont  $m$  et  $n_j$ , respectivement,  $j = 1, 2$ . En d'autres termes,  $Np_j/n_j \rightarrow 1$ ,  $j = 1, 2$ , et  $Np_1 p_2/m \rightarrow 1$ , lorsque  $N \rightarrow \infty$ . On obtient ainsi  $N_i$  comme estimateur de  $\text{Var}(N_i)$ .

**Exemple 2.** Supposons que les paires  $(p_{1i}, p_{2i})$ ,  $i = 1, \dots, N$ , sont indépendantes en ce sens que la distribution des probabilités  $p_{1i}$  est la même pour chaque valeur de  $p_{2i}$ . Alors,  $p_{12} = p_1 p_2$ ,  $S_3 = S_1 S_2$ ,  $S_4 = p_2 S_1$ ,  $S_5 = p_2 S_1$ . Par substitution dans l'équation du lemme 1, nous obtenons:

$$\text{Var}(N_i) = N \left( \frac{p_1 p_2}{1} - \frac{p_2}{1} - \frac{p_1}{1} - \frac{S_1 S_2}{p_2^2} + \frac{p_2^2}{S_1} + \frac{p_2^2}{S_2} \right)$$
$$= N \left( \frac{q_1 q_2}{p_1 p_2} - \text{cv}(p_{1i})^2 - \text{cv}(p_{2i})^2 \right),$$

où  $\text{cv}(p_{ji}) = (S_j - p_j^2)/p_j$  est le coefficient de variation des valeurs  $p_{ji}$ ,  $j = 1, 2$ . Evidemment,  $\text{Var}(N_i) \leq Nq_1 q_2/(p_1 p_2)$ . Une comparaison avec l'exemple 1 permet de constater que  $N_i$  est un estimateur conservatif de  $\text{Var}(N_i)$  (c'est-à-dire que  $N_i$  est asymptotiquement trop élevé) lorsque les  $p_{1i}$  sont indépendantes des  $p_{2i}$ . Autrement dit, étant donné les moyennes  $p_j$ ,  $j = 1, 2$ , on obtient la variance la plus élevée dans des conditions d'homogénéité. [Cela rappelle la variance du nombre de succès dans les schémas de Bernoulli avec probabilité de succès variable; voir Feller (1968, p. 230-231).] Une comparaison avec l'exemple 1 permet de constater que  $N_i$  est un estimateur conservatif de  $\text{Var}(N_i)$  (c'est-à-dire que  $N_i$  est asymptotiquement trop élevé) lorsque les paires  $(p_{1i}, p_{2i})$  sont indépendantes. Notons que la condition d'indépendance suppose que  $C = 0$ .

Si les probabilités ne sont pas indépendantes, il n'est pas sûr que l'estimateur classique soit conservatif. Il existe néanmoins un estimateur conservatif. On l'obtient en augmentant  $\text{Var}(N_i)$  d'une quantité qui peut être estimée en fonction des variables observables. Nous faisons la démonstration de la proposition générale suivante en annexe.

**Proposition 1.** Un estimateur conservatif de  $\text{Var}(N_i)$  est

$$V_2 = (n_1^2 n_2^2 + n_2^2 m n_1 + n_1^2 m n_2)/m^3,$$

où  $n_j = n_j - m$ ,  $j = 1, 2$ .



bas dans l'exemple 1, lorsque les probabilités de saisie sont homogènes, la variance asymptotique de  $N$  est  $\text{Var}(N) = Nq_1q_2/(p_1p_2)$ , où  $q_j = 1 - p_j$ ,  $j = 1, 2$ . On peut alors estimer  $\text{Var}(N)$  par  $V_1 = n_1n_2u_1u_2/m^3$  (Sekar et Deming 1949, p. 114-115).

Cet article a pour but d'examiner le bien-fondé de l'estimateur de variance  $V_1$  et de comparer le biais de cet estimateur avec celui de  $N$ . Nous avons choisi d'étudier  $V_1$  car il n'avait jamais été donné auparavant de vérifier si cet estimateur est efficace dans des conditions d'hétérogénéité, même lorsque  $N$  est convergent. Cela a pu être vérifié. De même, il était difficile de savoir dans quelles circonstances  $V_1$  produit des estimations excessives et peut, par conséquent, aboutir à des intervalles de confiance valables malgré le biais de  $N$ . Nous avons vu par la suite que cela était possible dans des circonstances particulières avec des intervalles unilatéraux.

Dans la section 2, nous calculons la variance asymptotique de  $N$ , lorsque  $N \rightarrow \infty$ , et déterminons un estimateur conservatif de cette variance (désigné par  $V_2$ ) dans des conditions d'hétérogénéité quelconque. Autrement dit,  $V_2$  produit une surestimation de la variance asymptotique réelle. De cette façon, on pourrait espérer compenser le biais généralement négatif de  $N$  par une surestimation de la variance et obtenir quand même des intervalles de confiance valables. Malheureusement, cela ne semble possible que lorsque le biais  $N$  de est faible ou que  $N$  est petit. Dans la section 3, nous examinons le bien-fondé de  $V_1$  dans des conditions d'hétérogénéité gaussienne et déterminons un estimateur  $V_3$ , lequel est conservatif dans ces conditions particulières d'hétérogénéité. La caractéristique gaussienne n'est pas indispensable pour les arguments; ce qui compte, c'est que les moments des paires  $(p_{1i}, p_{2i})$  concordent avec ceux d'une distribution gaussienne à deux variables. Il est alors facile d'étudier l'effet de la corrélation entre  $p_{1i}$  et  $p_{2i}$  sur l'estimation de la variance car la corrélation ne peut être exprimée qu'en fonction d'un paramètre, soit le coefficient de corrélation de moments ordinaire. Dans la section 4, nous comparons le biais de l'estimateur de variance à celui de  $N$  en nous servant de données empiriques sur l'enregistrement des cas de maladie professionnelle en Finlande.

## 2. BIAIS ET VARIANCE DANS DES CONDITIONS D'HÉTÉROGÉNÉITÉ

Définissons  $p_{jN}$  comme la probabilité moyenne de saisie au temps  $j$ ,  $j = 1, 2$ ; et posons  $p_{12N}$  comme la moyenne des produits  $p_{1i}p_{2i}$ ,  $i = 1, \dots, N$ . Alors,  $C_N = p_{12N} - p_{1N}p_{2N}$  est la covariance des paires  $(p_{1i}, p_{2i})$ . Supposons que les limites  $p_{jN} \rightarrow p_j$ ,  $j = 1, 2$ ;  $p_{12N} \rightarrow p_{12}$ , et  $C_N \rightarrow C$  existent. En conséquence,  $N/N \rightarrow p_1p_2/p_{12}$ , de sorte que  $N/N - 1 \rightarrow -C/p_{12}$ , lorsque  $N \rightarrow \infty$ . C'est ce qu'on appelle le biais asymptotique de l'estimateur classique dans des conditions d'hétérogénéité. Fait digne de mention, ce biais ne dépend que des deux premiers moments de la distribution des paires  $(p_{1i}, p_{2i})$ . Il est notoire (Sekar et Deming 1949, p. 105-106; Seber 1982, p. 86) que lorsque la covariance est nulle ( $C = 0$ ), l'estimateur classique est convergent; si  $C > 0$ ,  $N$  produit une sous-estimation et dans le cas contraire, il produit une surestimation. Comme nous l'avons indiqué plus haut, nous cherchons particulièrement à savoir si  $V_1$  est efficace lorsque  $p_{jN}$  varie selon les individus et que  $C$  est égal à 0.

Nous allons maintenant calculer la variance asymptotique de l'estimateur classique selon notre modèle général d'hétérogénéité. Notons qu'il n'existe pas de variance finie car il y a une probabilité réelle que  $m = 0$ . La "variance asymptotique" désigne donc ici la variance de la distribution limite plutôt que la limite de la variance lorsque  $N \rightarrow \infty$ .

**Lemme 1.** La variance asymptotique de  $N$  est

$$\text{Var}(N) = N \left\{ \frac{p_1^2 p_2^2}{p_{12}^2} - \frac{p_1^2 p_2}{p_{12}^2} - \frac{p_1 p_2^2}{p_{12}^2} - \frac{p_1^2}{p_{12}^2} S_1 - \frac{p_1^2}{p_{12}^2} S_2 - \frac{p_1^2 p_2^2}{p_{12}^2} S_3 + 2 \left( \frac{p_1^3}{p_{12}^3} S_4 + \frac{p_1^2 p_2^2}{p_{12}^3} S_5 \right) \right\},$$

# Estimation de la variance dans un système de double enregistrement pour des populations hétérogènes

JUHA M. ALHO<sup>1</sup>

## RÉSUMÉ

L'estimateur de système dual de la taille d'une population peut être fortement biaisé s'il y a hétérogénéité des probabilités de saisie. Dans cet article, nous étudions le biais de l'estimateur de variance corrigé dans des conditions d'hétérogénéité. Nous montrons que l'estimateur habituel est conservatif, c'est-à-dire qu'il produit des estimations excessives, lorsque les deux systèmes d'enregistrement sont corrélés négativement ou pas du tout ou encore lorsque il existe une corrélation positive mais faible. S'il existe une corrélation positive forte entre les deux systèmes, l'estimateur peut produire des sous-estimations. Nous proposons donc deux estimateurs, que nous comparons au premier. L'un est conservatif dans des conditions d'hétérogénéité quelconque, l'autre l'est dans des conditions d'hétérogénéité gaussienne. Nous appliquons ensuite ces estimateurs à des données sur les maladies professionnelles en Finlande.

MOTS CLÉS: Saisie-resaisie; système dual; hétérogénéité; maladies professionnelles.

## 1. INTRODUCTION

Supposons  $N$  individus dans une population fermée. Le problème consiste à estimer  $N$  au moyen d'un système de double enregistrement. Nous procédons à un échantillonnage double, selon lequel  $n_j$  individus sont prélevés (saisis) au temps  $j$ ,  $j = 1, 2$ . Soit  $m_i$  le nombre d'individus prélevés deux fois. Définissons les variables indicatrices  $u_{ji}$  et  $m_i$  pour  $i = 1, \dots, N$ , telles que  $u_{ji} = 1$ , si et seulement si l'individu  $i$  est prélevé au temps  $j$  seulement ( $j = 1, 2$ ), et  $m_i = 1$  si et seulement si l'individu  $i$  est prélevé aux deux occasions. Autrement,  $u_{ji}$  et  $m_i$  sont égales à zéro. Définissons  $n_{ji} = u_{ji} + m_i$  comme l'indicateur de saisie au temps  $j$ ,  $j = 1, 2$ . De plus, posons  $M_i = u_{1i} + u_{2i} + m_i$  comme la variable indiquant la saisie d'un individu à au moins une occasion. Enfin, désignons par  $p_{ji} = E[n_{ji}]$ ,  $j = 1, 2$ ; et  $p_{12i} = E[m_i]$  les probabilités de saisie individuelles. Posons par hypothèse que les probabilités se situent strictement entre zéro et un. Le fait que les probabilités peuvent varier d'un individu à l'autre indique qu'il peut y avoir hétérogénéité des probabilités de saisie dans la population. Nous terminons la définition du modèle de double enregistrement (ou de saisie-resaisie) en supposant que les saisies sont des événements indépendants pour chaque individu, ou  $p_{12i} = p_{1i}p_{2i}$ , et que les vecteurs multinomiaux

$$(u_{1i}, u_{2i}, m_i, 1 - M_i) \sim \text{Mult}(1, p_{1i}q_{2i}, p_{2i}q_{1i}, p_{1i}p_{2i}, 1 - \phi_i),$$

où  $q_{ji} = 1 - p_{ji}$ ,  $j = 1, 2$ , et  $\phi_i = p_{1i} + p_{2i} - p_{1i}p_{2i}$ , sont indépendants pour  $i = 1, \dots, N$ . On sait très bien que lorsque les probabilités de saisie ne varient pas d'un individu à l'autre, c.-à-d.  $p_{ji} = p_j$ ,  $j = 1, 2$ , l'estimateur du maximum de vraisemblance de  $N$  est  $\hat{N} = n_1n_2/m$  (ou, plus précisément, l'entier le plus près de cette valeur vers le bas; voir Feller 1968, p. 46). Cet estimateur classique peut être fortement biaisé dans des conditions d'hétérogénéité (Seber 1982, p. 565; Burnham et Overton 1979, tableau 4, p. 931-932). Comme on peut le voir plus

<sup>1</sup> Juha M. Alho, Institute for Environmental Studies et Department of Statistics, University of Illinois at Urbana-Champaign, 1101 W. Peabody Dr., Urbana IL, 61801, U.S.A.

Dans le dernier article de ce numéro, Norris et Paton exposent dans ses grandes lignes l'Enquête sociale générale du Canada, qui en est à sa cinquième année d'existence. Ils énumèrent succinctement les besoins en données et commentent les cinq sujets abordés annuellement dans l'enquête. Ils décrivent aussi la méthodologie de l'enquête ainsi que des expériences qui ont été faites avec la composition aléatoire. Leur analyse des taux de non-réponse enregistrés depuis les débuts de l'enquête a une portée pour d'autres enquêtes téléphoniques.

Le rédacteur en chef



## Dans ce numéro

Ce numéro de *Techniques d'enquête* renferme une série d'articles très diversifiée. Dans le premier article, Alho examine divers estimateurs de la variance de la taille d'une population estimée au moyen d'un système de double enregistrement. Il étudie le biais de l'estimateur habituel de la variance suivant l'hypothèse de l'hétérogénéité de la population, ce biais étant déterminé normalement selon l'hypothèse de probabilités de saisie homogènes. L'auteur propose deux autres méthodes produisant de meilleurs résultats que si on utilisait les estimations fondées sur les données d'un seul passage. Ils appuient leurs propos d'un exemple concernant l'enquête nationale sur les sorties d'hôpital aux États-Unis (National Health Discharge Survey).

Dans leur article, Lavrakas, Settersten et Maier font une analyse descriptive du problème de l'attrition des panels dans les enquêtes; ils se servent pour cela des données de deux enquêtes téléphoniques à composition aléatoire. Cet article instruit convenablement le lecteur dans quelques-unes des causes de l'attrition et propose des façons d'en atténuer les effets. Sutrachar, Dagum et Solomon définissent un test exact permettant de vérifier s'il existe un mouvement saisonnier stable significatif dans des séries temporelles caractérisées par une structure saisonnière stable, à part de quelques variations annuelles. Les hypothèses du test  $F$  de l'analyse de variance ordinaire utilisé dans la méthode de désaisonnalisation  $X-1-ARIMA$  ne sont pas vérifiées lorsqu'il y a autocorrélation des résidus. En revanche, le test exact tient compte de cette possibilité. Les auteurs comparent les deux tests pour plusieurs séries socio-économiques canadiennes.

L'échantillonnage par la méthode des quotas est caractérisé par l'absence de processus randomisé. Il faut donc recourir à une forme quelconque de modélisation pour construire des estimateurs. La méthode classique consiste en la modélisation de superpopulations et à cet égard, Deville ouvre de nouvelles voies intéressantes. Il propose notamment de modéliser le processus d'échantillonnage. Il établit aussi des comparaisons avec l'échantillonnage aléatoire. Bien que les enquêtes-ménages soient un heureux moyen de recueillir des données sur des populations humaines, elles ne sont pas conçues pour étudier les caractéristiques de populations humaines mobiles, par exemple les visiteurs de musées ou de parcs, les acheteurs, etc. Kalton décrit divers plans de sondage destinés à des enquêtes ayant pour objet des flux de populations humaines et il donne des exemples d'enquêtes de ce genre. Ces exemples montrent que le choix d'un plan de sondage est déterminé largement par les considérations relatives au travail sur le terrain.

Hidiroglou, Choudhry et Lavallée présentent un plan de sondage pour des enquêtes intra-annuelles permanentes auprès d'entreprises. Un plan de renouvellement est proposé afin que l'échantillon ne cesse d'être représentatif. À l'aide d'une étude empirique où sont reproduites certaines conditions d'enquête, les auteurs examinent les propriétés d'un certain nombre d'estimateurs de totaux utilisés avec ce plan de sondage. Stasny, Goel et Rumsey utilisent des modèles de régression pour établir des estimations régionales de la production de blé dans le cas où les sources de données sont non probablistes. Par une étude de simulation, il comparent les estimations obtenues par régression à celles calculées au moyen de deux estimateurs classiques: l'estimateur synthétique et l'estimateur direct. Ils comparent aussi trois méthodes de pondération qui visent à satisfaire les règles d'additivité.



# TECHNIQUES D'ENQUÊTE

Une revue de Statistique Canada  
Volume 17, numéro 2, décembre 1991

## TABLe DES MATIÈRES

Dans ce numéro .....	133
J.M. ALHO	
Estimation de la variance dans un système de double enregistrement pour des populations hétérogènes .....	135
P.S.R.S. RAO et I.M. SHIMIZU	
Combinaison d'estimations provenant d'enquêtes .....	143
P.J. LAVRAKAS, R.A. SETTERSTEN, Jr. et R.A. MAIER, Jr.	
Perte d'effectifs dans un panel obtenu par composition aléatoire dans deux enquêtes locales .....	155
B.C. SUTRADHAR, E.B. DAGUM et B. SOLOMON	
Test exact pour vérifier la présence d'un mouvement saisonnier stable et applications .....	167
J.-C. DEVILLE	
Une théorie des enquêtes par quotas .....	177
G. KALTON	
L'Echantillonnage des flux de populations humaines mobiles .....	197
M.A. HIDIROGLOU, G.H. CHOUDHRY et P. LAVALLÉE	
Méthodes d'échantillonnage et d'estimation pour des enquêtes infra-annuelles auprès des entreprises .....	211
E.A. STASNY, P.K. GOEL et D.J. RUMSEY	
Estimation de la production de blé par comité .....	229
D.A. NORRIS et D.G. PATON	
L'Enquête sociale générale canadienne: bilan des cinq premières années .....	245
Remerciements .....	261



# TECHNIQUES D'ENQUÊTE

## Une revue de Statistique Canada

La revue Techniques d'enquête est répertoriée dans The Survey Statistician et Statistical Theory and Methods Abstracts. On peut en trouver les références dans Current Index to Statistics, et Journal Contents in Qualitative Methods.

### COMITÉ DE DIRECTION

#### Président

G.J. Brackstone

#### Membres

B.N. Chinnappa  
G.J.C. Hole  
F. Mayda (Directeur de la production)  
R. Platek (Ancien président)  
M.P. Singh  
C. Patrick  
D. Roy  
M.P. Singh

### COMITÉ DE RÉDACTION

#### Rédacteur en chef

M.P. Singh, *Statistique Canada*

#### Rédacteurs associés

B. Afonja, *Nations Unies*  
D.R. Bellhouse, *U. of Western Ontario*  
D. Binder, *Statistique Canada*  
E.B. Dagum, *Statistique Canada*  
J.-C. Deville, *INSEE*  
D. Drew, *Statistique Canada*  
W.A. Fuller, *Iowa State University*  
J.F. Gentleman, *Statistique Canada*  
M. Gonzalez, *U.S. Office of Management and Budget*  
R.M. Groves, *U.S. Bureau of the Census*  
D. Holt, *University of Southampton*  
G. Kalton, *University of Michigan*  
J.N.K. Rao, *Carleton University*  
L.-P. Rivest, *Université Laval*  
D.B. Rubin, *Harvard University*  
I. Sande, *Bell Communications Research, U.S.A.*  
C.E. Särndal, *Université de Montréal*  
W.L. Schaible, *U.S. Bureau of Labor Statistics*  
F.J. Scheuren, *U.S. Internal Revenue Service*  
C.M. Suchindran, *University of North Carolina*  
J. Waksberg, *Westat Inc.*  
K.M. Wolter, *A.C. Nielsen, U.S.A.*

#### Rédacteurs adjoints

J. Gambino, L. Mach, H. Mantel et A. Thèberge, *Statistique Canada*

### POLITIQUE DE RÉDACTION

La revue Techniques d'enquête publie des articles sur les divers aspects des méthodes statistiques qui intéressent un organisme statistique comme, par exemple, les problèmes de conception découlant de contraintes d'ordre pratique, l'utilisation de différentes sources de données et de méthodes de collecte, les erreurs dans les enquêtes, l'évaluation des enquêtes, la recherche sur les méthodes d'enquête, l'analyse des séries chronologiques, la désaisonnalisation, les études démographiques, l'intégration de données statistiques, les méthodes d'estimation et d'analyse de données et le développement de systèmes généralisés. Une importance particulière est accordée à l'élaboration et à l'évaluation de méthodes qui ont été utilisées pour la collecte de données ou appliquées à des données réelles. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de Statistique Canada.

#### Présentation de textes pour la revue

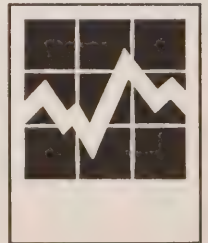
La revue Techniques d'enquête est publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à en faire parvenir le texte au rédacteur en chef, M. M.P. Singh, Division des méthodes d'enquêtes sociales, Statistique Canada, Tunney's Pasture, Ottawa (Ontario), Canada K1A 0T6. Prière d'envoyer quatre exemplaires dactylographiés selon les directives présentées dans la revue. Ces exemplaires ne seront pas retournés à l'auteur.

#### Abonnement

Le prix de la revue Techniques d'enquête (catalogue n° 12-001) est de 35 \$ par année au Canada, 42 \$ (É.-U.) aux États-Unis, et de 49 \$ (É.-U.) par année à l'étranger. Prière de faire parvenir votre demande d'abonnement à Section des ventes des publications, Statistique Canada, Ottawa (Ontario), Canada K1A 0T6. Un prix réduit est offert aux membres de l'American Statistical Association, l'Association Internationale de Statisticiens d'Enquête et la Société Statistique du Canada.

# Techniques d'enquête

Statistique Canada  
Division des méthodes d'enquêtes sociales



Une revue de Statistique Canada  
Décembre 1991 Volume 17 Numéro 2

Publication autorisée par le ministre  
responsable de Statistique Canada

© Ministre de l'Industrie, des Sciences  
et de la Technologie, 1991

Tous droits réservés. Il est interdit de reproduire ou de  
transmettre le contenu de la présente publication, sous quelque  
forme ou par quelque moyen que ce soit, enregistré ou non,  
support magnétique, reproduction électronique, mécanique,  
photographique, ou autre, ou de l'emmagasiner dans un système  
de recouvrement, sans l'autorisation écrite préalable du Chef,  
Services aux auteurs, Division des publications, Statistique  
Canada, Ottawa, Ontario, Canada K1A 0T6.

Décembre 1991

Prix : Canada : 35 \$

États-Unis : 42 \$ US

Autres pays : 49 \$ US

Catalogue 12-001

ISSN 0714-0045

Ottawa







# Techniques d'enquête

Une revue de Statistique Canada

Décembre 1991 Volume 17 Numéro 2

Catalogue 12-001

374000036









JUN 8 1994



